One rule to rule them all

Representations of learned rules for categorization in mouse prefrontal cortex

Sandra Reinert

Graduate School of Systemic Neurosciences
LMU Munich

Dissertation der
Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

October, 2022

Supervisor:
Dr. Pieter Goltstein
Max Planck Institute of Neurobiology

First Reviewer:	Prof. Dr. Mark Hübener
Second Reviewer:	Prof. Dr. Christian Leibold
Date of Submission:	29.10.21
Date of Defense:	19.01.22

Contents

Abstract

Learning to form categories of objects and experiences is a fundamental skill for humans and many animals. Grouping stimuli based on their appearance or on their behavioral relevance, and storing these groups in our memory, helps us to react quickly to novel stimuli. Particularly by learning rules for categorization, we can direct our attention to features that are informative for sorting stimuli into categories. Such rule-based categorization allows us to flexibly adapt to a change in contexts and, therefore, a change in rules. In humans and non-human primates, prefrontal cortex has been identified as one of the key brain areas for learning rule-based categorization. However, the neuronal mechanisms, within prefrontal cortex and larger networks of brain areas, that underlie category learning remain largely unknown.

In this dissertation, I investigate rule-based category learning in mice and characterize a representation of learned categories in prefrontal cortex. I conducted two studies to achieve this: first, I established optimal training parameters for head-restrained operant conditioning in mice, and second, I recorded neuronal activity in prefrontal cortex throughout learning in a mouse model of rule-based category learning.

In the first study, I contrasted two nutritional restriction regimes in their effect on animal welfare and learning performance in an operant conditioning task. Learning paradigms often involve restricting mice in their *ad libitum* consumption of food or water, in order to motivate them to participate in operant conditioning paradigms. Then, typically soy milk or water is given as a reward upon upon correct behavioral choices. I compared how food- and water restriction regimes affected the welfare of mice and their performance in a visual stimulus discrimination task. I found that, overall, animal welfare was largely unaffected by either regime, while water-restricted mice on average showed mildly higher discomfort levels. Food- and water-restricted mice achieved similar plateau performances in the learning task, but water-restricted mice learned significantly faster. In summary, in this study I determined optimal training parameters for operant conditioning experiments, while keeping mouse discomfort to a minimum (Goltstein et al., 2018a).

In the second study, I established a mouse model of rule-based category learning and characterized the emergence of a category representation in prefrontal cortex. From humans and non-human primate research, it is known that rule-based category learning relies on selective attention to category-defining features. While there are indications that rodents are also able to selectively attend to specific features or modalities, it was unclear whether mice could learn rule-based categorization. I first showed that mice learned to categorize, generalized to novel stimuli and successfully performed a rule-switch. Based on these behavioral results, I concluded that mice had learned rules for categorization. By chronically recording neuronal activity in prefrontal cortex, I was able to characterize the responsiveness of individual neurons throughout category learning. I found that, after learning, neurons in prefrontal cortex showed category-selective responses. This neuronal category representation emerged gradually, as mice were learning the visual categories, and was specific to the learned rule. Lastly, I discovered that part of this category representation generalized across different operant behaviors and reward contingencies. This finding points toward a semantic component of the prefrontal cortical representation that is independent of learned operant behavior. Thus, this study showed for the first time that mice can learn rules for categorization and that single cell responses in mouse prefrontal cortex reflect learned categories. Characterizing the emergence of such responses over the course of learning expanded upon important findings from human and non-human primate category learning research (Reinert et al., 2021).

Taken together, the studies in this thesis establish the mouse as a model system for investigating rule-based category learning and identify category-selective neurons in prefrontal cortex as a key component of the underlying neuronal circuitry.

1 | Introduction

A young monkey in the Kenyan savanna is on guard duty today. He has to watch out for potential threats through either leopards, eagles or pythons. For each of these predators he has learned a different alarm call from his conspecifics. When his troop hears a 'leopard call', they will know that they should hide up on a tree. In contrast, in response to an 'eagle call' the group will hide on the ground in a bush or after a 'python call' rather frantically search the ground. This monkey's ability to identify these three kinds of predators, to distinguish them from animals that do not pose a threat and to communicate so precisely, is fundamentally important to the survival of his troop.

Because the vervet monkeys reliably detect those predator species in vastly different contexts, but do not produce alarm calls to non-threatening animals like pigeons or geese, their behavior is an example for categorization (Seyfarth et al., 1980). Further, young vervet monkeys do not exactly know yet, when to produce a specific alarm call. Infants emit an 'eagle call', both when they spot an eagle, or a pigeon. Likewise, when hearing a call from a conspecific, they first watch the adults in the group for their reaction. In the first four years of their lives, they learn to produce the alarm calls in response to the appropriate predators and to react with the appropriate behavior to the alarm calls (Seyfarth and Cheney, 1986).

Such categorization and learning of categories are remarkable features of human and animal behavior that influence almost every of our daily-live decisions. How we form categories and what underlying changes happen in the brain, to date remains an unsolved question.

In this dissertation, I will first give an overview of category learning research from a behavioral, neuronal and computational point and discuss what benefits a mouse model can bring to our understanding of category learning processes in the brain. I will then address this question by presenting two studies. In the first study, I optimized operant conditioning parameters for a category learning paradigm in mice, and in the second study, I followed individual neurons in the mouse prefrontal cortex throughout such category learning using chronic two-photon calcium imaging.

1.1 Category learning

What is a category? A category is a set of objects, stimuli or experiences grouped together based on similar perceptual features or a similar required reaction. The vervet monkeys have learned three categories of predators, each eliciting a different alarm call and behavioral reaction (Seyfarth et al., 1980). Categorization of sensory inputs or experiences is seen universally, from arthropods to mammals. A cricket, for example, uses sound cues in its environment to determine whether the sound source is a conspecific or rather a predator. Conspecifics call at a frequency of 4-5 kHz, whereas predators, mostly bats, produce ultrasound at 25-80 kHz. Somewhere along the continuous spectrum of sound frequencies (between 13 and 16 kHz) crickets sharply transition their decision from attraction to a conspecific to escape from a predator. Thus, there are two categories of sound frequencies eliciting different behavior (Wyttenbach et al., 1996).

However, there likely is a difference between the categorization behavior of vervet monkeys and crickets. The crickets did not have to learn these categories, they show a form of innate categorization. The vervet monkeys, on the other hand, learn the predator categories from the members of their troop during development and shape their alarm calls and responses accordingly (Seyfarth and Cheney, 1986).

Innate categories can be observed throughout the animal kingdom, indicating that categorization provides advantages that are crucial for an animals survival: Categories allow to react fast to novel stimuli without the need to memorize all individual items. However, innate categories are likely genetically determined and hence restricted to one process or one modality. Hence, that benefit does not extend to other situations or other decisions that need to be made. In order to flexibly apply the advantage of categories to whatever might become relevant, one needs a system that is able to learn novel categories.

Category learning has traditionally been studied in humans. Consequently, theories for mechanisms of category learning, and also early models for their neural basis, have been largely based on results from human behavioral investigations. Therefore, I will first discuss how category learning is typically assessed and then describe key experiments and theories, before giving an overview of categorization in non-mammalian and mammalian animals.

1.1.1 Human category learning

1.1.1.1 Category learning and memory

Category learning research is a part of memory research. Breaking down the formation of a category into steps, the learning process likely starts with individual experiences (inputs) that become linked to a certain behavioral reaction. These experiences will be remembered and can be recalled for future decisions. As a second step, similar inputs (or inputs leading to a similar reaction) will be associated with each other. Presumably through these two steps, forming individual memories and associating them with each other, a category is learned. Thus, learning and memorizing a new category relies on the ability of the brain to encode, store and retrieve information. These processes are not unique to category learning, but integral to memory in general. Therefore the research into memory systems has guided also the investigations and theories of category learning.

At the end of the 19th and beginning of the 20th century several psychologists and neuroscientists transitioned from viewing memory as a single function to describing different 'kinds' of memory. Distinctions were made, for example, between 'memory' and 'habit' (James, 1913) or implicit and explicit recall (McDougall, 1923; for review see Squire, 2004). In 1904, Richard Semon coined the term 'engram' as the neural substrate of stored information, and this idea of a memory trace in the brain pushed the search for the underlying structure of memory forward (Lashley, 1950; for review see Rolls, 2000; Hübener and Bonhoeffer, 2010). But only from the second half of the 20th century on, significant progress was made in unifying views on different memory systems and understanding mechanisms of memory formation.

A prominent starting point of the search for the engram as a substrate of memory was the case of patient H.M. In 1953, patient H.M. underwent surgery to relieve him from epileptic seizures. In this surgery, large parts of both temporal lobes, including both hippocampi and parahippocampal regions, were removed. Unfortunately, as a side effect of this surgery, patient H.M. struggled with the formation and recall of memories. Specifically, while he could remember childhood memories and could still learn motor skills, like complex mirror drawing tasks, he could not form memories of any events that happened after the surgery (Penfield and Milner, 1958; Corkin, 1968). From case studies of amnesiacs like patient H.M. (Milner et al., 1968; Warrington and Weiskrantz, 1968) and increasing work in animal models (Hirsh, 1974; O'Keefe and Nadel, 1978; Squire and Zola-Morgan, 1991), the idea of multiple memory systems gained popularity, with the dichotomy of declarative, or explicit, vs non-declarative, implicit, memory at its core. Subsequently, different brain areas were identified as the key players for declarative and specific types of non-declarative memory (Tulving, 1985; Squire, 1987, 2004). Declarative memory was further divided into episodic and semantic memory, referring to events and facts, respectively. Non-declarative memory was used as an umbrella term for memory types like procedural memory and classical conditioning.

The story of understanding category learning parallels this search for the engram. The popular view of category learning has changed from theories proposing one system to multiple systems theories, as we will see when exploring several models of human categorization.

1.1.1.2 Assessing human category learning

In this thesis, I will focus on research of learning of novel categories and disregard the work on categorization performance of highly trained experts. Such expert categorization, like the (in)famous chicken sexers (i.e. the profession of determining the sex of young chickens), is acquired over years of training and very specific to the trained categories. Therefore it is not readily comparable to studies of category learning that probe performance on newly acquired categories within few trials or sessions (Ashby and Maddox, 2005). For the ease of explanation of experiments and theories, I will refer to items that are categorized as stimuli. However, all the discussed theories can, in principle, also be applied to objects and abstract concepts or constructs.

Categorization performance is typically assessed as the fraction of stimuli that are assigned correctly to the category label, i.e. a percentage of correct trials, or the derived discriminability (*d'*) of the two categories. These metrics can be calculated in an ongoing fashion during training and give an indication of the progress in learning to categorize stimuli. However, aside from categorizing stimuli that have been trained, a hallmark of category learning is to use the learned category information and apply it to novel stimuli that have not yet been encountered. This process is called generalization. Generalization can also be evaluated with performance and discriminability metrics, by exclusively considering trials in which novel stimuli were tested. Another measure is the accuracy of the learned category boundary (i.e. the separation between two categories in the feature space) in comparison to the true category boundary. The learned boundary is hereby typically estimated through modelling multiple possible boundaries and determining which of those optimally predicts the subject's category decisions (Maddox and Ashby, 1993; Smith et al., 2010).

Human category learning has been tested both in basic and clinical research in a wide variety of task designs. Categories contained few or many stimuli that varied across one or multiple stimulus dimensions. Further, the stimuli in each category were either normally distributed or followed a different distribution, and the category boundary could be linear or non-linear. Category learning was also evaluated for a variety of sensory modalities, although by far the largest proportion of experiments used visual stimuli.

Paradigms for human category learning have been sorted into three major groups of task designs: prototype distortion tasks, rule-based and information-integration tasks. Prototype distortion tasks are paradigms in which presented images are stochastically distorted around a fixed prototype. Rule-based and information-integration tasks differ in the number of stimulus dimensions that are relevant to the category boundary. In rule-based categorization, categories can be formed using one informative stimulus dimension, whereas tasks that require integration of two or more stimulus dimensions are called information-integration tasks. In a rule-based paradigm, the category boundary can be described with a verbalizable rule, whereas the information-integration task cannot be solved with an easily verbalizable rule (Ashby et al., 1998).

One of the most prominent examples for a rule-based task design is the Wisconsin Card Sorting Test (WCST; Robinson et al., 1980; Heaton and Pendleton, 1981) first described by Esta Berg (1948). The test requires a subject to sort playing cards according to rules that are uninstructed and change between blocks of trials. The rule that a subject needs to apply is always one-dimensional and easy to verbalize. This test is widely used as a diagnostic tool to probe for working memory, attention and behavioral flexibility, but requires the rapid categorization of stimuli based on a one-dimensional rule.

In summary, the metrics to assess category learning are applied in largely the same fashion, but the task designs that test categorization vary in several parameters. In the following sections, I will give an overview of the behavioral results in various categorization tasks, how they have been interpreted to support theories of human category learning and how the task design affects observed categorization.

1.1.1.3 Theories proposing a single category learning system

Early on, psychologists tried to conceptualize how humans categorize stimuli. The dominant idea was that there is one mechanism, i.e. one system, that is applied to every categorization problem. Several theories were formulated, and behavioral experiments were designed to test predictions of such theories (for review see Ashby and Maddox, 2005).

The classical view of category learning assumes that a category is a representation of a group as a whole, defined by a set of features that are both necessary and sufficient for classification (Hull, 1920; Medin and Smith, 1981). While the classical view of categories existed in more or less the same form since Aristotle, more modern theories emerged in the second half of the 20th century and mostly built on the classical view. I will first consider the prototype theory. According to the prototype theory of concepts, the formation of a category involves three elements. The first element is a prototype representation, which can either be the average of all encountered stimuli or one typical stimulus falling into that category. Second, the theory relies on a way to calculate the similarity of any stimulus to the prototype stimulus and third, there needs to be a criterion of similarity to pass in order to determine category membership (Hampton, 1995). Thus, according to the prototype theory, a learned category consists of a memorized prototype item and a threshold criterion for similarity. Every encountered stimulus will be compared to the category prototype and, if the similarity exceeds the threshold, considered as part of the category.

The prototype theory predicts that the difficulty of categorization increases with the distance to the stored prototype and that prototype stimuli hold a unique position within the categories (Posner and Keele, 1968). Experimentally, these predictions were mostly tested using prototype distortion tasks. First evidence in favor of this theory, published by Posner et al. (1967), showed that the rate of learning of categorization decreased with the amount of distortion from the prototype. Further, the prototype (average) of all trained images was more likely to be recognized than other new patterns within a category, even if it was never presented. This result was in line with the second prediction of the theory (Posner and Keele, 1968).

The later emerging exemplar theory (sometimes referred to as context theory) postulates that during category learning individual exemplars (stimuli) are memorized rather than one prototype. The stored exemplars define the characteristic feature space of the category. When a novel stimulus is encountered, specific exemplars will be recalled, depending on their similarity to the novel stimulus. The novel item will then be compared to the retrieved exemplars and attributed to the category with largest sum of similarities (Medin and Schaffer, 1978; Nosofsky, 1986). In contrast to the prototype theory, no abstraction of a category prototype is needed to sort an item into a category. Rather, the specific retrieval of exemplar information is sufficient. The exemplar theory therefore solves categorization problems with non-linear boundaries and dependencies in features better than the prototype theory. Medin and Schaffer (1978) created categories where prototype and exemplar predictions would clash. Specifically, stimuli could be sorted into categories based on similarity to the experienced exemplars or to the category prototype. The performance of human subjects rather depended on similarity to experienced stimuli rather than the distance to the category prototype. Therefore, the authors concluded that the exemplar theory better explained categorization performance.

The most recent model for category learning that describes a single system is the decision bound model (Ashby and Gott, 1988; Maddox and Ashby, 1993). It postulates that categories are learned by dividing a perceptual space into regions that are associated with different responses. The partition between the regions is referred to as the decision bound, i.e. the learned category boundary. A critical difference between this theory and the prototype or exemplar theories is that in the decision process no similarity metric (to the average or individual exemplars) is calculated. Rather, a response label can be directly retrieved because the representation of the perceptual space is separated into regions associated with a response. Maddox and Ashby (1993) found that this decision bound model predicted subjects' performances in categorization tasks with non-linear boundaries better than the exemplar or the prototype theory, likely because it better co-varied with perceptual imperfections (noise) in the categorization decision.

In summary, these three proposed mechanisms of category learning step-by-step improved in predicting human data, by addressing the shortcomings of older theories like differences in category structure and suboptimal behavior by the subjects. Importantly, each of these models was designed with a specific categorization task at hand and, with some exceptions, was mainly tested on that type of task. The prototype theory was tested and provided a good explanation for prototype distortion tasks (Reed, 1972), the exemplar model was typically tested with tasks requiring the integration of two stimulus dimensions to categorize small sets of stimuli, and the decision bound model was typically applied to tasks with two large, normally distributed categories. However, a systematic comparison of theories across task designs could help to better understand the underlying mechanisms. In the following sections, I will contrast the behavior observed in the various tasks with specific attention to similarities and differences in the attempt to form a coherent picture of human category learning.

1.1.1.4 Category learning is influenced by task type

Based on the observation that different human category learning studies supported different theories of category learning, researchers tried to vary individual task parameters in order to determine their effect on categorization performance. Hereby, evidence accumulated that the performance depended on many aspects, such as learning stage (Smith and Minda, 1998), category size (Homa et al., 1981; Minda and Smith, 2001; Katz et al., 2002), variance within categories (Smith et al., 1997; Blair and Homa, 2003) or whether a rule is easy or difficult to verbalize (Ashby and Maddox, 2005).

I want to specifically highlight one aspect: the number of stimulus features that determine the category identity, i.e. the dimensionality of the category boundary. Shepard et al. (1961) found that the more stimulus dimensions needed to be integrated, the more difficult the task became and the less subjects were able to generalize to novel stimuli. The dimensionality of the category boundary is also the major distinction between rule-based and information-integration tasks. The two task designs are, by now, commonly contrasted using visual stimuli, typically oriented gratings, that vary in two different visual features, like their orientation and spatial frequency (Fig. 1.1; Ashby et al., 1999; Maddox et al., 2003; Smith et al., 2010; Smith et al., 2012).

Human subjects categorize visual stimuli in rule-based tasks faster and to a better plateau performance than identical stimuli in information-integration tasks (Smith et al., 2010). Smith et al. created a rule-based (RB) model, that only took one stimulus dimension into account, and an information-integration (II) model, that linearly combined the two dimensions, and fit both to the decisions of every subject. In a rule-based task, the performance of the majority of subjects was best captured by the RB model, whereas the II model best explained categorization of most subjects in the information-integration task. However, for some subjects, the RB model best predicted the performance even in information-integration tasks. This phenomenon is referred to as rule-bias and highlights a tendency to prioritize one stimulus dimension. Hence, humans show a preference for rule-based strategies, even though that impairs their performance in some tasks (Smith et al., 2010; Vermaercke et al., 2014). The observed performance advantage in rule-based tasks, but also the rule-bias, indicate that humans separately analyze stimulus dimensions and direct selective attention towards informative dimensions (Smith et al., 2010).

Aside from selective attention, three further factors distinguished rule-based and information-integration tasks because they specifically affected one of them: novel stimuli, feedback timing and retinotopic location of stimulus presentation. First, when human subjects were trained on a rule-based task, they could generalize the learned rule to novel stimuli. This ability was impaired in information-integration tasks, especially if the tested stimuli were far from the trained category space (Maddox et al., 2005; Casale et al., 2012). Second, delaying the timing of the feedback that a subject is given during the category learning only impaired the performance in information-integration task and not in rule-based tasks (Maddox et al., 2003). In the absence of feedback, humans tended to apply rule-based strategies even if they were suboptimal to solve the task. Third, the performance of human subjects in a visual information-integration task was impaired when the stimuli were presented in a spatial, retinotopic position that was different from the trained location of the visual field, in contrast to performance in a rule-based category learning task (Rosedahl et al., 2018).

In summary, human categorization performance strongly depended on the testing conditions. This was particularly emphasized by the contrasting effect of selective attention, feedback and stimulus position on performance and generalization in rule-based and information-integration tasks. Together, these observations indicate that humans learn categories with more than one, uniformly underlying mechanism.

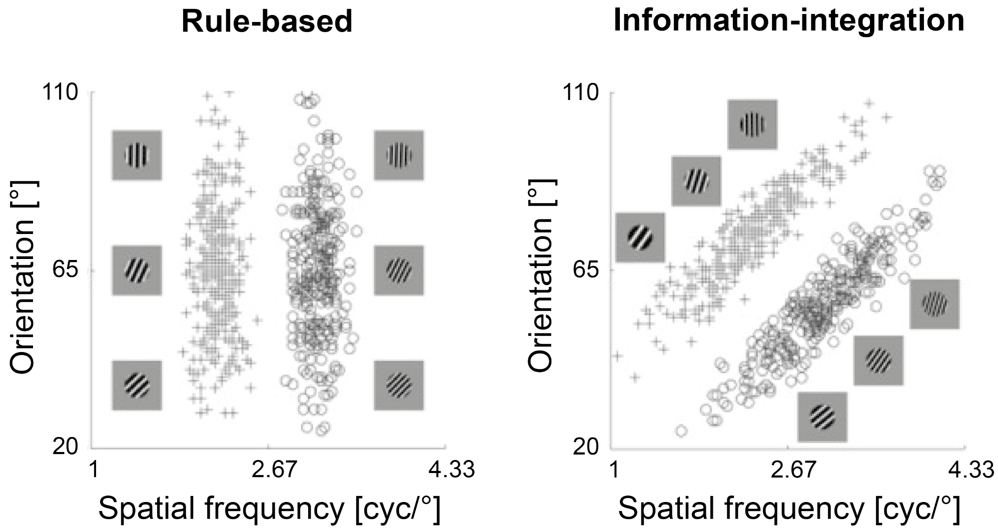


Figure 1.1: **Rule-based and information-integration categories.** Visual stimuli with two features, orientation and spatial frequency of bars, can be selected to create either a rule-based (left) or an information-integration (right) category structure. Each panel depicts a stimulus space from which stimuli of category A (crosses) or B (circles) are randomly drawn. The distribution of stimuli in each category and the distance between categories, hence the difficulty, can be kept identical, so that the only difference between the two tasks is the dimensionality of the category boundary. In the rule-based task, a one-dimensional boundary defines the categories, whereas in the information-integration task the boundary is two-dimensional (Reprinted and adapted from Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., Spiering, B., Beran, M. J., Church, B. A., Ashby, F. G., & Grace, R. C. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2355–2369, Copyright (2012), with permission from Elsevier).

1.1.1.5 Multiple category learning systems theory

Both, the influence of task design on human category learning and the fact that none of the single system theories (see 1.1.1.3) were able to explain this effect, motivated the theory that category learning is mediated by multiple systems. Since memory research most prominently distinguishes between declarative/explicit memory and non-declarative/implicit memory, category research has attempted to align theories of category learning with those memory systems (Allen and Brooks, 1991; Ashby and O'Brien, 2005; Smith et al., 2012).

Within the framework of implicit and explicit systems, implicit category learning is considered a slow process of associating responses to individual stimuli. Because, in this case, subjects do not have conscious access to the reasoning for their category decisions, this process is akin to procedural memory formation. This contrasts with explicit category learning, a conscious process of analyzing stimulus features, relying on selective attention and often resulting in a simple, verbalizable rule for categorization. Explicit category learning therefore shows close resemblance to a declarative, semantic memory formation process.

This distinction between an explicit and an implicit category learning system is able to explain the observed differences between rule-based and information-integration tasks (see 1.1.1.4).

Even though these two tasks test categorization of the same stimuli with the same perceptual difficulty, they are likely solved using different strategies. The rule-based task engages an explicit category learning process, supported by selective attention and the formation of a verbalizable rule (Maddox et al., 2005; Casale et al., 2012). In contrast, the information-integration task involves a more implicit, procedural process of categorization that relies on immediate feedback and is dependent on the stimulus position (Ashby et al., 1999; Maddox et al., 2003; Rosedahl et al., 2018). Taken together, these observed effects can be explained by the multiple systems theory of category learning.

Aside from categorization in adults, also investigations into changes of category learning strategies during development support the multiple systems theory. Nelson (Nelson, 1984) tested the influence of the subjects' age on the categorization performance in a rule-based and an information-integration task. The study examined five year old children and compared their performance to ten year old children. Strikingly, half of the 5 year old children failed to learn the rule-based categorization, whereas the ten year old children learned both tasks equally well. This difference points to a tendency of younger children to use implicit category learning. Older children are able to identify and selectively attend to informative stimulus dimensions, like adults do. Such a change during development suggests that the implicit category learning system is the older, more fundamental category learning strategy.

In summary, the multiple systems theory of category learning can explain results from human category learning studies better than single systems theories. Explicit tasks are learned faster and to a higher performance than implicit categorization tasks. Both the tendency towards the explicit category learning system in adults and the observation that this tendency arises during development, encourage the detailed investigation into the brain structures that might form the basis for the different category learning systems (see 1.2). Importantly, the discussed rule-based and information-integration tasks can also be tested in several other species, which allows for a comparison between human and animal categorization. Such a comparison can give us a better understanding of conserved mechanisms as well as potentially uniquely human features.

1.1.2 Category learning in animals

„*The proper study of mankind is man* is a popular quote, but it was written by a poet, not a scientist. The history of science offers opposing testimony [...] if you want to build a scientific understanding of the evolution and meaning of intelligence then you must study animals[...]. The eventual payoff will indeed be an understanding of people as well as beasts."
quoted from Staddon (2016) by Zentall et al. (2008).

1.1.2.1 Categorization is universal in the animal kingdom

Category learning is a complex cognitive process in humans, and has traditionally been researched in an anthropocentric fashion (Zentall et al., 2008). Theories like the prototype, exemplar or decision bound theory were developed based on human categorization experiments and were exclusively tested using human data. On the other hand, examples of categorization can be found universally through the animal kingdom, from the categorical sound perception of crickets (Wyttenbach et al., 1996) to the specific calls vervet monkeys use to communicate threat by different types of predators (Seyfarth et al., 1980).

While advances in neuroimaging strongly contribute to human category research in healthy subjects and patients, animal models like rodents and primates offer yet a different set of methods to study category learning and underlying changes in the brain. By investigating categorization and category learning in different animals, we can potentially identify conserved mechanisms and discover what features of category learning are uniquely mammalian or even human. Key to the benefits of category research in animals is developing appropriate tasks and model systems that allow for a comparison to human data and theories (Zola-Morgan et al., 1983; Squire, 2004;

Zentall et al., 2008; Smith et al., 2012). In the following sections, I will discuss category research in non-mammalian (see 1.1.2.2) and mammalian (see 1.1.2.3) animal models, specifically focusing on the comparability and generalizability of the results.

1.1.2.2 Non-mammalian categorization

When it comes to eating and being eaten, i.e. detection of prey or predators, researchers have found evidence of categorization almost universally, whether studied in arthropods, fish or birds. But is non-mammalian categorization exclusively innate or can we observe category learning? If so, can we detect commonalities and differences in category learning systems compared to humans?

In arthropods, crickets show the ability to categorize prey sounds from predator sounds (Wyttenbach et al., 1996). Experiments on jumping spiders have shown that they can detect their preferred prey, even if the experimenters just showed abstract versions, i.e. lines with conserved dimensions and angles, to the spiders (Dolev and Nelson, 2014). However, most of the observed categorization behaviors appear to be innate rather than learned, with a notable exception being honey bees learning visual categorization (Benard et al., 2006).

In contrast to the mostly innate categorization in arthropods, birds show remarkable skills in learning novel categories. Early on, Herrnstein and Loveland (Herrnstein and Loveland, 1964) discovered that pigeons can learn to categorize pictures containing humans and contrast them to pictures without humans. Category learning in pigeons was not specific to classifying humans (Herrnstein and Loveland, 1964; Aust and Huber, 2001; Yamazaki et al., 2007) , but could also be learned for categories like trees, bodies of water and individual persons (Herrnstein et al., 1976) or even painter styles (Watanabe et al., 1995). Crucially, pigeons could apply the learned response to novel images of the same category, i.e. generalize.

Smith et al. (2011) tested pigeons on the same information-integration categories and rule-based categories as they used to test humans (Smith et al., 2010), allowing for a direct comparison of categorization behavior. The study showed that pigeons were able to learn both tasks. However, in contrast to the findings of the human experiments, pigeons did not show a learning rate or performance advantage for the rule-based task. When modeling the performance in the information-integration task, a diagonal decision bound best explained the behavior of all pigeons, indicating that they efficiently learned the categories and were not biased to one-dimensional, rule-biased categorization. Taken together, these results suggest that pigeons solve rule-based categorization with a similar strategy as the information-integration task, rather than showing evidence for different category learning systems. This interpretation has led to the hypothesis that pigeons lack the explicit category learning system or at least show a strong dominance of their implicit category learning system (Smith et al., 2012).

On the other hand, results from Yamazaki et al. (2007) contrast this hypothesis. The researchers presented pigeons with stimuli either binocularly or monocularly to either eye. This study found that stimuli presented to the left eye were learned faster initially, but the right eye was more robust to scrambling of small features. Because in the pigeon each eye nearly exclusively projects to the respective contralateral hemisphere, these results indicated that the right hemisphere (left eye) rather categorized based on familiarity in an implicit learning fashion, whereas the left hemisphere (right eye) categorized based on category-defining features, i.e. more analytical. This observation of lateralization conflicts with the hypothesis that pigeons only have an implicit category learning system.

It is important to keep in mind that the observed difference between human and pigeon category learning (Smith et al., 2011; Smith et al., 2012) could also have been due to a anthropocentric way of testing for explicit categorization, i.e. by using stimulus dimensions that are perceptually separable for humans but perhaps not for pigeons.

So while categorization and the ability to form new categories, are clearly not uniquely mammalian (or human) features, comparing the performance in explicit and implicit categorization tasks has pointed to differences in strategies or potentially even category learning systems between non-mammalian species and humans. More research is needed to characterize avian category learning strategies and systems to determine if birds also employ implicit and explicit category learning despite very different brain structures (see 1.2.1) or where and when the explicit system has evolved.

1.1.2.3 Mammalian categorization

Is explicit category learning a uniquely human feature, or where, and how often, has the explicit, analytical system evolved? Studying mammalian animal models in category learning tasks can help to address this question and potentially contribute to the understanding of prerequisites, applications and advantages of an explicit category system.

The animals that are evolutionarily closest to the human are without a doubt non-human primates. Hence, it was natural to investigate category learning in monkeys, most prominently in rhesus macaques and capuchin monkeys. While the spectrum of tested tasks and category structures resembles the variety used in human category learning research, the distinction between an implicit and an explicit category learning system has also been addressed. Similar to humans, both macaques and capuchin monkeys learned rule-based categorization faster and to a higher end performance compared to the information-integration task (Smith et al., 2010; Smith et al., 2012). These results show that explicit categorization is not a uniquely human ability and that, contrary to early theories (Shepard et al., 1961; Ashby et al., 1998), it is not necessarily linked to language. Non-human primates clearly possess a toolkit for dimensional analysis and rule formation similar to humans, even if in a potentially rudimentary form.

The finding that an explicit category learning system is not uniquely human encouraged the investigation of other mammals. In the history of cognition research, rodents remained largely underexplored, likely in part because of the difficulty to train them on complex tasks because rodents often showed suboptimal behavioral strategies or because of differences in sensory processing compared to humans or primates (Churchland and Kiani, 2016; Nakajima and Schmitt, 2019). However, in the question of the evolution of an explicit categorization system, studying rodents could provide important information, especially when comparing their results to pigeons and primates. Rats have been trained in rule-based and information-integration tasks, similar to the other species explored so far (Vermaercke et al., 2014; Broschard et al., 2019b; Broschard et al., 2020), but the question whether rats have an explicit category learning system remains unresolved. Similar to pigeons, rats did not show a learning rate or performance advantage of a rule-based task compared to an information-integration task (Vermaercke et al., 2014; Broschard et al., 2019b; Broschard et al., 2020). Modeling the categorization strategy of rats in both tasks, however, revealed that in the rule-based task roughly half of the animals were best fit by a unidimensional strategy and the other half by a two-dimensional strategy. Further, all animals in the rule-based task differentially weighed the two stimulus dimensions, pointing to selective attention towards one dimension even in the absence of a fully rule-based strategy (Broschard et al., 2019b).

Category research in mice has shown that they are able to form arbitrary stimulus categories (Runyan et al., 2017; Xin et al., 2019; Zhong et al., 2019), recognize the category of three-dimensional objects (Creighton et al., 2019) and even categorize paintings according to different painters (Watanabe, 2017). However, no experiments have specifically addressed explicit versus implicit category learning in mice, even though mice have been increasingly tested in higher cognitive abilities (Rikhye et al., 2018; Zempeltzi et al., 2020). Category learning has either been demonstrated along a single stimulus dimension (Runyan et al., 2017; Xin et al., 2019; Zhong et al., 2019), or on highly dimensional objects (Watanabe, 2017; Creighton et al., 2019) Neither allows for a controlled assessment of selective attention, dimensionalization or categorization strategy.

In summary, mammals like primates and rodents learn categories in a way that resembles humans, showing features of explicit categorization like selective attention and rule formation, even though they lack language. It is possible that the observed similarities and differences in categorization behavior in pigeons, rodents, non-human primates and humans are due to similarities and differences in the underlying brain structures and their learning mechanisms.

Studying the brain mechanisms of category learning in both non-mammalian and mammalian animals can help us understand human cognitive processes better.

1.2 Neuroscience of category learning

One of the central aims of neuroscience is to understand the human brain and with it the mechanisms for human cognition. In the previous section, I have explored how universal categorization is in the animal kingdom. However, since brain structures have changed with evolution, likely also the brain mechanisms underlying categorization and category learning differ between species depending on their evolutionary distance. One obvious example is the distinction between an organism with an innate categorization mechanism and any organism that is able to learn novel categories. Since the majority of arthropod studies describe innate rather than learned categorization, I will not consider the arthropod brain for this overview. I will briefly touch upon category learning in the avian brain and, in the later sections, focus on brains that are structurally most similar to the human brain, mammalian brains.

1.2.1 Category learning in the avian brain

Category research in birds has largely used visual categories, hence I will focus on the avian visual system and not consider other sensory modalities. In the avian brain, visual information is processed by two parallel pathways, the thalamofugal pathway, corresponding to the mammalian geniculocortical tract, and the tectofugal pathway, comparable to the mammalian extrageniculocortical system. In contrast to the mammalian visual paths, the avian tectofugal pathway conveys information about the frontal visual field and the thalamofugal pathway relays information from the lateral visual field. In visual association areas like the nidopallium frontolaterale (NFL), the mesopallium ventrolaterale (MVL) and the visual Wulst, inputs from both pathways converge. These visual association areas are reciprocally connected with the nidopallium caudolaterale (NCL), an area that resembles the mammalian prefrontal cortex in connectivity (Kröner and Güntürkün, 1999) and functional properties (Güntürkün, 1997; Kirsch et al., 2009).

First evidence for categorical representations in the avian brain was found in the NCL (Kirsch et al., 2009). In pigeons trained in a visual categorization paradigm, individual neurons displayed activity patterns following the category identity, i.e. the functional meaning, of stimuli rather than responding to individual stimuli based on their visual features. These category-selective responses strengthened with learning and shifted temporally from being rather reward-related to stimulus-related.

Investigating neurons in the visual association areas NFL and MVL revealed that neuronal populations encoded basic visual stimulus categories, like pictorial stimuli versus gratings, even in animals that had not undergone any conditioning (Koenen et al., 2016; Azizi et al., 2019). In contrast to NCL, these areas hence formed perceptual categories based on visual statistics of the environment rather than functional meaning. During category learning, such representations could be combined with learned relevance in NCL populations (Güntürkün et al., 2018).

In summary, in the avian brain that lacks cortex (the mammalian structure most category learning research focusses on) learned category representations can be detected in the nidopallium caudolaterale. This observation and the excellent categorization performance (see 1.1.2.2) with such different brain structures, encourage comparing category learning circuits across animal classes. Such a comparison could help understand how different systems enable similar behaviors, potentially by engaging similar circuit mechanisms of learning (Güntürkün et al., 2018).

1.2.2 Category learning in the mammalian brain

When looking for neuronal underpinnings of category learning in the mammalian brain, I will not consider sensory epithelia and stages of processing before the primary sensory cortices. Even though each sensory input pathway involves several earlier stages and there is evidence of experience dependent changes in subcortical sensory processing (Jaepel et al., 2017), it is likely that early stages, such as retinae and sensory thalamus, provide rather stable information about the environment (Gilbert and Wiesel, 1992). So far, changes in neuronal activation induced by category learning have been investigated in cortical sensory processing areas and, cortical and subcortical, association areas, but not earlier subcortical sensory processing areas, like thalamic nuclei. Therefore I will focus the following sections on brain areas that are primary sensory cortices and downstream areas.

1.2.2.1 Sensory cortices

Primary sensory cortices, for instance primary visual or auditory cortex, receive and represent information about our environment in an organized fashion, for example based on retinotopy (Dräger, 1975) or tonotopy (Romani et al., 1982). They are the first cortical processing stage that encode sensory information about the stimuli that we categorize.

One can envisage two scenarios on how sensory cortices could be involved in category learning. First, information could 'just' be relayed to association areas that would combine features and extract relevant information for the categorization process. Thus, the role of sensory cortices in category learning would be similar to the role of the retinae, conveying visual information irrespective of any category learning or memory. However, sensory areas could also have a more direct role in category learning by, for instance, improving the distinction of stimuli at the borders of a category space and decreasing the distinction within a category space. Such adaptations during learning could hence improve categorization performance.

Experimental results indicate that the involvement of sensory cortices depends on the task design. On the one hand, human neuropsychological and neuroimaging data point towards an involvement of visual cortex in prototype distortion tasks that are thought to rely on perceptual learning (Reber et al., 1998). In line with this, higher sensory areas like the inferior temporal cortex (ITC) in primates show improved selectivity towards a feature that is relevant to a learned categorical distinction and some category-selective responses for the learned categories (Sigala and Logothetis, 2002; Freedman et al., 2003; Kiani et al., 2007; Brincat et al., 2018). In mice that have learned sound frequency categories, auditory cortex populations show higher responses to stimuli near the category boundary (Xin et al., 2019). Activity patterns in gerbils trained on such a frequency categorization become more invariant to the individual stimulus frequency, hence more similarly represent stimuli of the same category (Ohl et al., 2001). On the other hand, rule-based and information-integration tasks do not elicit specific activation in visual cortical areas in humans (Nomura et al., 2007). Likewise, electrophysiological recordings in macaque lower visual areas, V4 and middle temporal (MT), showed that neurons in these areas predominantly represent the sensory information of a stimulus, irrespective of a learned categorical distinction (Brincat et al., 2018).

Across species, in some tasks sensory areas show changes in sensory encoding with category learning. These changes improve discrimination between different categories or weaken within category discrimination, together resulting in better categorization. These results argue that sensory areas are to some extent involved in the learning of categories and do not just relay information to downstream areas. However, the magnitude of the detected effects is overall subtle compared to higher association areas like parietal or prefrontal cortices (see 1.2.2.2,1.2.3.3).

1.2.2.2 Parietal cortex

Parietal cortex, specifically posterior parietal cortex (PPC), has traditionally been seen as a higher order sensory processing area that receives visual, auditory and somatosensory inputs (for review see Lynch, 1980). This view was challenged, prominently by Mountcastle et al. (1975), finding robust activation of monkey PPC through motor action. Over several decades of research, human and monkey PPC (for a review on homology see Orban et al., 2006) has been shown to be involved in sensorimotor integration (Duhamel et al., 1992; for review see Freedman and Ibos, 2018), decision making and motor planning (Shadlen and Newsome, 2001) and selective attention (Yantis et al., 2002; Yantis and Serences, 2003; Behrmann et al., 2004). Neuroimaging studies have found transient increases in PPC activity whenever subjects shifted their attention from one spatial location to another (Yantis et al., 2002; Yantis and Serences, 2003), from one object-feature to another (Liu et al., 2003) or from one sensory modality to another (Shomstein and Yantis, 2004). Recently, several studies have found similar coding in rodent PPC (Harvey et al., 2012; Whitlock, 2014; Runyan et al., 2017), although drawing homologies of specific subregions of PPC between species remains difficult (Whitlock, 2014).

Because of its implication in selective attention and decision making, posterior parietal cortex is by now considered an association area (Fitzgerald et al., 2011; Whitlock, 2017) and likely also plays a role in category learning. PPC could link perceptual information from sensory areas to learned motor responses and hence hold a category representation (Seger and Miller, 2010). Indeed, human parietal areas showed specific activation during visual categorization tasks (Aizenstein et al., 2000; Vogels et al., 2002). In line with the human results, in monkey lateral intraparietal area (LIP), an area within primate PPC, category-selective neuronal responses were identified after learning, indicating that PPC represents visual information based on behavioral relevance (Freedman and Assad, 2006). Supporting these results, *in vivo* two-photon imaging and optogenetic interventions in mouse PPC during an auditory category learning task found a causal contribution of PPC neuronal activity to categorization of new stimuli and reassigning stimuli according to a new category boundary (Zhong et al., 2019). However, although PPC neurons stably represented the learned categories, its activity was not necessary for categorizing well-learned stimuli (Zhong et al., 2019).

In summary, the posterior parietal cortex is likely causally involved in category learning and holds a representation of learned categories. Whether the mechanism of causal contribution is more direct, that is, through forming a category memory and comparing individual stimuli to that, or more indirect, as in providing more basic functions that are necessary for the categorization like selective attention or high level feature integration, needs to be investigated further. Similarly, how the observed category representations are computed and learned, i.e. bottom-up or top-down, remains unclear, although there are indications that category representations do not rely on feedback projections from prefrontal areas (Swaminathan and Freedman, 2012).

1.2.2.3 Basal ganglia

The mammalian basal ganglia consist of the striatum, the globus pallidus, the ventral pallidum, the substantia nigra, and the subthalamic nucleus. The striatum is further subdivided into the dorsal striatum, containing the caudate nucleus and putamen, and the ventral striatum, comprised of the nucleus accumbens and olfactory tubercle. Traditionally, basal ganglia were implicated with voluntary motor control and action selection (Graybiel, 2005; Seger, 2008), based on the most obvious symptoms in patients with a basal ganglia dysfunction like Parkinson's disease or Huntington's disease (Jankovic and Tolosa, 2007). However, studies have also noted a role of the basal ganglia in procedural learning (Knopman and Nissen, 1991; Pascual-Leone et al., 1993; Knowlton et al., 1996) and set shifting, that is in tasks that involve a change in task-rule (Hayes et al., 1998; Monchi et al., 2006). These findings have put the basal ganglia, and specifically the (dorsal) striatum, at the center of mechanisms of stimulus-response learning (Barnes et al., 2005), i.e. the formation of stimulus-outcome associations, and therefore basal ganglia might also be crucial for category learning.

Indeed, there is evidence for a role of basal ganglia in category learning. Patients with Parkinson's disease show an impairment in information-integration categorization (Cools et al., 1984; Maddox and Filoteo, 2001), and to a lesser extent rule-based categorization (Price et al., 2009), compared to healthy subjects. Further, several neuroimaging studies found the striatum to be activated during category learning (Reber et al., 1998; Seger and Cincotta, 2002; Vogels et al., 2002; Lim et al., 2019). Notably, different categorization tasks elicited activation in different regions of striatum, indicating that the involvement of striatum in category learning is heterogeneous. Subjects performing a task similar to the WCST showed activation in the head of the caudate nucleus (Rao, Bobholz, et al., 1997). This area also shows severe damage in Parkinson's patients, congruent with their observed deficit in set shifting (Cools et al., 1984; Hayes et al., 1998). In contrast, the tail of the caudate nucleus showed learning-related activity changes in subjects performing an information-integration category learning task (Poldrack et al., 1999).

Electrophysiological recordings in the caudate nucleus of macaques found neurons selective to a learned rule for categorization (Merchant et al., 1997; Muhammad et al., 2006) or categories in a prototype distortion task (Antzoulatos and Miller, 2011). In the latter study, the striatum formed category-selective responses early in the training phase as the animal was learning a novel category, faster than prefrontal cortex. These responses likely correspond to striatal stimulus-response mapping or exemplar learning. As the category size increased, category selectivity in striatum decreased, presumably because mapping individual stimuli grew inefficient, i.e. abstraction of category knowledge became necessary.

In rodents, the striatum is involved in procedural learning (McDonald and White, 1994; Jog et al., 1999; Costa et al., 2004), strategy selection (Whishaw et al., 1987; Packard and McGaugh, 1992) and set shifting (Ragozzino, 2007; Lindgren et al., 2013). Like in humans and non-human primates the striatum is functionally heterogeneous (Devan et al., 1999; Pistell et al., 2009). A potential involvement of rodent striatum in category learning has not yet been assessed. Although it is likely that rodent striatum provides a similar toolkit as found in primates, precise homologies between caudate nucleus in primates and dorsal striatal structures in rodents remain unclear (Balsters et al., 2020).

Taken together, human and primate basal ganglia show an involvement in both explicit and implicit category learning mechanisms. This is likely due to their role in fast learning of stimulus-outcome mappings that underlies early learning in both systems. Further, the functional heterogeneity of striatal substructures implicates that they are part of separate, larger networks of brain areas that coordinate category learning within the two systems.

1.2.2.4 Hippocampus

As already discussed (see 1.1.1.1), the hippocampus is a brain area central to memory formation, especially to spatial memory (Morris et al., 1982) and episodic memory (Penfield and Milner, 1958). A prominent theory on memory consolidation proposes that memories are formed in hippocampus and sequentially, through cortical projections, are externalized to cortical networks for long-term storage (McClelland et al., 1995; Squire and Alvarez, 1995). Both, the involvement in declarative memory systems and promising mechanistic models of memory formation and consolidation suggest a role of hippocampus also in category learning.

Early evidence was mainly gathered in patients suffering from medial temporal lobe amnesia. Several studies have observed that patients show specific deficits in category learning in tasks involving few stimuli, when there was no one-dimensional rule, i.e. information-integration tasks or rule-based tasks with complex rules (Knowlton et al., 1994; Rickard and Grafman, 1998; Reed and Squire, 1999). This deficit manifested mainly later in training (Knowlton et al., 1994). Interestingly, amnesiacs did not show any impairment compared to healthy subjects in information-integration tasks involving large sets of stimuli (Filoteo et al., 2001). Possibly, when encountering few exemplars of categories, hippocampus forms memories of these. As soon as the number of exemplars exceeds the capacity of memorizing individual stimuli, procedural learning systems involving the basal ganglia dominate the category learning process (Gluck et al., 1996; Reed and Squire, 1999).

Neuroimaging studies support this theory, finding involvement of hippocampus in both information-integration and rule-based tasks (Little et al., 2006; Nomura et al., 2007; Seger et al., 2011; Mack et al., 2016; Bowman and Zeithamova, 2018), specifically early in training (Little et al., 2006). When subjects were asked how they solved the categorization, trials in which they reported that they actively remembered a specific stimulus to make a decision, hippocampus was activated (Seger et al., 2011). In a categorization task with a one- or two-dimensional rule, hippocampal activation in response to an object changed when the category identity of the object changed. This activation reflected the attentional weighing of (ir)relevant stimulus features (Mack et al., 2016). A theoretical model, 'EpCon' (Mack et al., 2018), proposes that hippocampus forms categories from encoding individual exemplars, i.e. episodes, and subsequent pattern completion and pattern separation mechanisms.

A category representation in individual hippocampal neurons was described by Hampson and colleagues (Hampson et al., 2004), who identified category-selective cells in the hippocampus of primates categorizing clip-art images, i.e. multidimensional categories. The study further demonstrated that those neurons generalized their responses to novel images according to features that the images had in common with the learned category. Kim et al. (2018) confirmed in rats that hippocampus plays a causal role in categorization by bilaterally inactivating the hippocampi using the GABA agonist Muscimol. During the inactivation, animals that were trained in a categorization task with few exemplars showed impaired performance on both trained and novel stimuli.

In summary, findings across species show that the hippocampus plays a role in learning categories. However, the potentially diverse effects of category structure and time in training on the involvement of hippocampus (that have been observed in human neuropsychological data) have not yet been explored in animal models. Since intact hippocampal function is not necessary for every category learning process, it is unlikely to be the one categorization area, but rather part of a network of brain areas. While there are theoretical models of how hippocampus can acquire and update representations of categories, circuit models that aim to integrate hippocampal function with other brain areas are lacking.

1.2.3 The role of prefrontal cortex in category learning

Prefrontal cortex is a higher association area that is implicated with diverse functions like working memory, attention and behavioral inhibition. Both, structural and functional considerations of prefrontal cortex have been refined over decades of research. Further, it is still debated to what extent prefrontal cortex can be compared across mammalian species (Uylings et al., 2003; Seamans et al., 2008; Carlén, 2017). Therefore, I will first provide an overview of structural definitions and functional considerations separately, before discussing the role of prefrontal cortex in category learning.

1.2.3.1 Structural considerations of prefrontal cortex

In its widest definition, prefrontal cortex is the cerebral cortex of the frontal part of the frontal lobe (Brodmann, 1909). Beyond that, across-species definitions of prefrontal cortex are still debated. Early definitions of prefrontal cortex were based on cytoarchitecture, specifically on the presence of a granular Layer IV (Jacobsen, 1935). This definition was used to establish homology between human and non-human primate brain areas. However, non-primates, e.g. rodents, do not have a granular cortex in the frontal lobe, prompting a debate about the existence of any homologue to prefrontal cortex in rodents (Uylings et al., 2003; Seamans et al., 2008; Carlén, 2017). Another definition considers prefrontal cortex as the projection zone of mediodorsal (MD) thalamus (Rose and Woolsey, 1948), based on the idea that cortical differentiation may be regulated by thalamocortical connections (Rakic, 1988). Despite differences in cytoarchitecture, this definition enabled drawing homologies between humans, non-human primates and also non-primates. Later however, more refined experimental methods have shown that prefrontal areas also receive projections from other thalamic nuclei and that MD thalamus also projects to other cortical areas such as premotor or parietal cortex (Giguere and Goldman-Rakic, 1988; Morán and Reinoso-Suárez, 1988). This lead to a more refined version of the projection-based definition: all areas with preferential connectivity to MD thalamus over other thalamic nuclei (Uylings and Eden, 1991) are considered prefrontal cortical areas.

Historically, prefrontal cortex attracted the attention of neuroscientists, as the area of largest growth in humans (phylogenetically: Brodmann, 1909; for review see Fuster, 2002; but also see Semendeferi et al., 2002 and ontogenetically: Jernigan and Tallal, 1990; Pfefferbaum et al., 1994; Sowell et al., 1999). Further, prefrontal cortex is considered one of the most interconnected cortical areas (Miller and Cohen, 2001; Ährlund-Richter et al., 2019) giving it the reputation of a 'hub', managing other brain areas. With respect to subcortical areas, prefrontal cortex shows connections to thalamus (Rose and Woolsey, 1948), the basal ganglia (Alexander et al., 1986) and hippocampus (Hoesen, 1982; Goldman-Rakic et al., 1984). Its cortical connections involve several sensory areas (Pandya and Yeterian, 1990), parietal cortex (Jones and Powell, 1970; Goldman-Rakic and Schwartz, 1982), contralateral prefrontal cortex (Goldman-Rakic and Schwartz, 1982) and premotor areas (Lu et al., 1994). Notably, most of these connections are reciprocal.

Because the phylogenetic and ontogenetic development of PFC parallels the development of higher cognitive function (Gibson, 1991) and the interconnectivity of prefrontal areas with sensory, association and motor areas, prefrontal cortex has always been linked to a role in higher cognitive functions, before there was functional evidence. When discussing prefrontal cortex function and its role in category learning, I will follow the definition of PFC based on preferential connectivity with MD thalamus. With this definition I will try to form a coherent picture of prefrontal cortex in category learning that is also supported by functional homologies (see 1.2.3.2). Nevertheless, it is important to keep in mind that prefrontal cortex in every species encompasses several subregions, showing structural and functional variation within and across species (Fig. 1.2; Fuster, 2002; Carlén, 2017; Merre et al., 2021).

1.2.3.2 Functional considerations of prefrontal cortex

Historically, there were only diffuse accounts of functions of the frontal lobes obtained through lesion studies in humans and animal models (for review see Jacobsen, 1928). That was likely due to the lack of specificity in the location of the lesion, very diverse tests of cognitive and motor functions and, therefore, low reproducibility of results. The frontal lobes were suggested to play a role in learned motor behaviors, behavioral inhibition and attention. Patients with specific types of frontal lobe lesions showed an impairment in the Stroop task (developed by Stroop, 1935). In this test, subjects are presented with words that spell out a color, written in a coherent or incoherent font color, and are required to name the font color. Difficulty in performing the Stroop task indicate a deficit in assigning selective attention to a relevant feature in order to resolve an interference (Perret, 1974; Vendrell et al., 1995).

A notable point in the study of prefrontal cortex was the discovery of 'memory cells'. Electrophysiological recordings in macaques trained in a delayed response task reported cells in prefrontal cortex that showed activity spanning the delay period between a stimulus and the behavioral response (Fuster and Alexander, 1971; Goldman-Rakic, 1995). These first recordings were supported by neuroimaging studies (Jonides et al., 1993; Cohen et al., 1994), suggesting a role for PFC in working memory. In addition to neurons with sustained activity during delay periods, Quintana and Fuster (1999) discovered neurons whose activity ramped up in preparation of a specific motor response. Together, these findings indicated a role for PFC in working memory, action selection and planning and therefore confirmed its importance in attentional processes and goal-directed behavior.

While memory cells and preparatory activity were specific examples of PFC encoding

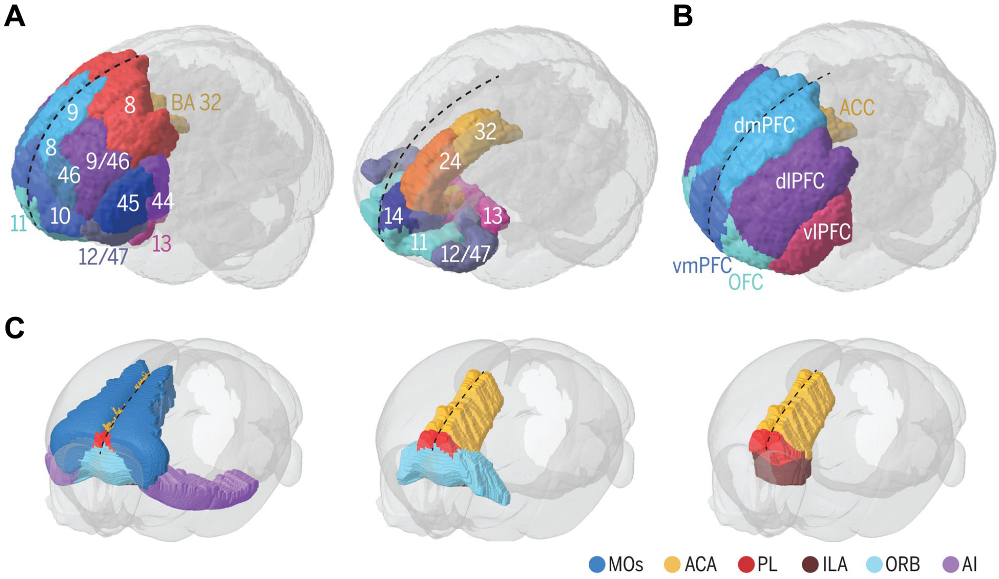


Figure 1.2: **Human and mouse prefrontal cortex.** (A) Schematic of Brodmann areas of human prefrontal cortex, based on the architectonic data from the Brainnetome Atlas (Fan et al., 2016). (B) Schematic of commonly used functional subdivisions of human prefrontal cortex. dmPFC: dorsomedial prefrontal cortex; vmPFC: ventromedial prefrontal cortex; vlPFC: ventrolateral prefrontal cortex; OFC: oribital frontal cortex. (C) Schematic of mouse prefrontal cortex subdivisions based on the Allen Brain Atlas (Lein et al., 2006). MOs: secondary motor area; ACA: anterior cingulate area; PL: prelimbic area; ILA: infralimbic area, ORB: orbital area; AI: agranular insular area. Dashed black line: sagittal midline. (From Carlén, M. (2017). What constitutes the prefrontal cortex? *Science*, *358*(6362), 478–482. Reprinted with permission from AAAS.).

parameters that are crucial to a task, is was soon shown to represent task-rules and task-relevant parameters in general. Rats with a lesioned prefrontal cortex showed an impaired ability to adapt to a change in task-rules (Winocur and Eskes, 1998; Ragozzino et al., 1999). Further, in cross-modal decision making tasks, a set of neurons in primate and rodent prefrontal cortex fired in correlation with a learned combination of stimuli, a task-rule (Fuster et al., 2000; Rikhye et al., 2018). Wallis et al. (2001) discovered that such a single cell rule-encoding also extends to abstract rules like ‘same vs. different’. However, not only on the level of individual neurons, but also on the population level, prefrontal cortex flexibly represents task-relevant parameters, such as context, stimuli and reward (Rao, Rainer, et al., 1997; Rainer et al., 1998; Mante et al., 2013; Lak et al., 2020).

Aside from PFCs involvement in goal-directed behavior in the form of working memory, selective attention and action planning, it has also been implicated in forms of long-term memory (for review see Simons and Spiers, 2003). Patients with frontal lobe lesions showed impaired memory of the temporal order of encountered objects despite otherwise normal object recall (Shimamura et al., 1990). Neuroimaging in humans (Tulving et al., 1996; Rugg et al., 2002) and studies in primates and mice (Xiang and Brown, 2004; Ye et al., 2016; Kitamura et al., 2017), further supported the role of PFC in the storage and recollection of episodic memory, potentially through an interaction with hippocampus (Simons and Spiers, 2003).

In search for a unifying principle of PFC function with respect to its involvement in goal-directed behavior, different and partially overlapping models have been developed (Duncan, 2001; Fuster, 2003). Fuster (2003) hypothesized that the main function of prefrontal cortex was to temporally organize behavior in order to reach a goal. This function would be served by temporal integration of perception and action through its role in coordinating attention, working memory, motor planning and monitoring of proprioceptive feedback. The author hereby proposed a double role for PFC: on the one hand, it would serve in temporally organizing cognitive functions and, on the other hand, it would represent relevant parameters as a rule or a concept and store them as part of long-term memory. Duncan (2001) formed an adaptive coding model of PFC. The key feature in this theory is the suggested adaptability of individual neurons in their coding.

Prefrontal neurons that are highly flexible in what they represent get assigned to encode relevant parameters whenever a subject is in a specific task context. As such, the prefrontal cortex operates as a global workspace that determines relevant parameters and in turn shapes the encoding of these in other brain regions, likely through attentional modulation. However, this model leaves open questions about the stability of such PFC representations and therefore its potential role in long-term memory.

1.2.3.3 Involvement of prefrontal cortex in category learning

The central role of prefrontal cortex in goal-directed behavior suggested that it also plays an important role in category learning. Already Milner (1963), and later several others, showed that patients with frontal lobe lesions were unable to perform in the WCST, hence had difficulties to form categories based on rules. Even though patients learned the first rule, they were unable to update their sorting behavior in response to rule changes (Milner, 1963; Robinson et al., 1980). These results are supported by human neuroimaging during rule-based tasks (Volz et al., 1997; Konishi et al., 1999) and lesion studies in monkeys (Dias et al., 1996; Rossi et al., 2007). In contrast, learning implicit, information-integration categories was not impaired in frontal lobe patients (Knowlton et al., 1996) and fMRI studies found less activation in frontal areas during implicit categorization compared to explicit categorization conditions (Seger and Cincotta, 2002).

Broschard et al. (2021) specifically compared the categorization behavior in information-integration tasks and rule-based tasks in rats with prefrontal cortex lesions. Hereby, they found that lesioned rats were impaired in learning the rule-based categorization compared to unlesioned control animals, but not impaired in the information-integration task. Together, the results indicated that prefrontal cortex plays a specific role in explicit category learning, in line with the earlier discussed importance in flexible rule-formation and attentional processes (see 1.2.3.2).

Electrophysiological recordings in monkeys added more detail to the contribution of prefrontal cortex to category learning. Freedman et al. (2001) identified neurons in primate prefrontal cortex showing category-specific activation, after the animals had been trained to categorize images of cats and dogs. Further studies discovered that, when animals were trained on different rules for categorization, the categories were represented by largely independent groups of such category-selective neurons (Roy et al., 2010). With learning, this category representation in prefrontal cortex developed slower than in striatum (Antzoulatos and Miller, 2011; Villagrasa et al., 2018) and posterior parietal cortex (Swaminathan and Freedman, 2012), but faster than in inferior temporal cortex (Freedman et al., 2003). In addition to the temporal sequence of activation of these brain areas, differences in the level of categorical abstraction of the represented stimuli were found, with the highest level of abstraction in prefrontal cortex (Brincat et al., 2018). Although in human category research a large focus is on comparing implicit and explicit category learning with respect to the involvement of different brain areas, in primates or rodents prefrontal cortex activity has not yet been characterized in information-integration category learning tasks. Hence, a specific role of prefrontal cortex in explicit category learning processes over implicit ones is yet to be reproduced in animal models in order to better understand the contribution of PFC neurons to the category learning systems in the brain.

A recent study in humans (Mack et al., 2020) indicated that PFC contributed to category learning by compressing high-dimensional inputs to relevant features and using that to modulate activity in other brain areas. The researchers found through fMRI recordings that the dimensionality of neural activation patterns in prefrontal cortex correlated with the dimensionality of the category learning task. The authors hypothesized that prefrontal cortex reduces the dimensionality of information to the relevant parameters for categorization. With such neural compression, PFC instructs selective attention by modulating neuronal coding in other brain regions, like medial temporal lobe areas (Mack et al., 2016). However, how neural compression to relevant features is achieved and how it would be reflected in the activity of individual neurons in PFC remains unclear.

Across species, PFC has been identified as a major player in category learning. Specifically, rule-based, explicit category learning relying on selective attention, abstraction and generalization processes strongly involves prefrontal areas. However, there is a disconnect in theoretical models that are based on human data, focusing on comparing different category learning systems, and the studies that are largely done in primate and rodent models, focusing on prefrontal cortex activity during explicit categorization. The promising finding that rodents, like humans, require PFC for rule-based category learning and not information-integration learning suggests that across species there might be a common way for the brain to solve common problems, so that this gap could be bridged.

1.2.4 Circuit models of category learning

1.2.4.1 Multiple systems for category learning in the brain

Similar to the theories of category learning behavior, there is a debate about how the brain implements category learning. With the behavioral data, I described the division between single systems theories and multiple systems theories. On an implementation level, this distinction could be further sorted into three possible scenarios. One hypothesis is that one brain area is the category learning area, irrespective of task or category aspects. A second possibility is that one network of interconnected brain structures implements one algorithmic mechanism of category learning. Both would fall in the group of single system theories. A third option is that the brain does not have one uniform way of learning categories but rather involves different brain structures depending on the underlying categorization problem. This multiple systems theory for the brain suggests that the network of brain areas that is best suited for a specific categorization problem will be recruited to solve the task.

At the core of single system theories for category learning is the assumption that one mechanism (or algorithm) is used to learn categories. Most models of prototype and exemplar theories (see 1.1.1.3) are agnostic to whether these algorithms are implemented by one brain area or distributed across many (Reilly et al., 1982; Kruschke, 1992). Crucial for this distinction is the answer to 'What determines a category learning area?'. Depending on the definition, an area that holds category-selective neurons, like PFC (see 1.2.3), could be seen as the category learning area, whereas a region that improves sensory discrimination near a category boundary or represents individual stimuli, such as sensory cortices and striatum, respectively (see 1.2.2), could be classified as not being category learning areas. Thus, the distinction between the first and the second scenario, one area or a network of areas, largely depends on the magnitude of effects of category learning. I will rather consider brain areas that show changes in neuronal activity in response to category learning as relevant regions, hence focus away from the possibility of the one category learning area.

A more popular view is that a network of brain areas together implements one mechanism for category learning. Reilly et al. (1982) break down the prototype and exemplar theory into individual algorithmic steps – 1) memory of individual exemplars and 2) fine-tuning their weights, i.e learning a threshold of similarity to each exemplar – and propose that cell assemblies in a network of brain areas could serve stages of this algorithm. The ALCOVE (Kruschke, 1992) and SUSTAIN (Love et al., 2004) models extend this algorithm with a weighting mechanism for relevant exemplar dimensions, i.e. selective attention, and a clustering mechanism by which exemplars are only memorized if they get falsely categorized at first. All three models are focused on improving the fit to human behavioral data. They break down categorization as how a given network of neurons could solve category learning problems, but do not include suggestions on which brain areas could implement such computations; data from neuropsychological studies and recordings are not considered.

The third group of theories for category learning propose that multiple systems coordinate category learning that are implemented by separate networks of brain regions. Roughly at the same time, two similar models were constructed: ATRIUM (Erickson and Kruschke, 1998) and COVIS (Ashby et al., 1998). Both contain two independent systems that learn in parallel and compete for the category decision. ATRIUM is an extension of the ALCOVE exemplar model

(Kruschke, 1992) that adds a rule learning system to the exemplar learning. The ATRIUM model could better explain rapid rule learning, which ALCOVE did not capture, and fit generalization performance more accurately. In contrast to ATRIUM, COVIS was constructed with data from neuropsychological and fMRI studies as a basis. Ashby et al. (1998) sorted findings of several studies based on the aspects of the tasks, category structures and further context, in order to determine general functional principles of individual brain areas. COVIS brings many such findings to a common denominator by proposing a prefrontal-based explicit system and a striatal-based procedural learning system.

Taken together, the debate whether a single or multiple systems implement category learning in the brain is not resolved yet (Nosofsky and Kruschke, 2002). Single system models lack concrete suggestions for brain areas that provide the proposed algorithms. COVIS on the other hand considers results from neuropsychological and neuroimaging studies. I will discuss this model in detail and present experimental support for its assumptions and predictions in the following section.

1.2.4.2 COVIS

One specific formulation of a theory for multiple systems for human category learning was developed by Ashby and colleagues (Ashby et al., 1998). The theory, COVIS (competition between verbal and implicit systems), was formed based on the multiple systems theory for memory (Tulving, 1985; see 1.1.1.1) and constrained by neuropsychological data. It proposes two category learning systems in the brain that closely relate to semantic memory and procedural memory systems. On the one side, a verbal, explicit, hypothesis testing system and on the other side an implicit, procedural learning system. In a task, these systems work in parallel, computing category decisions, but are competing for determining the behavioral output. The authors propose that, initially, both systems learn to form categories. Early in learning, the verbal system dominates the category decision, likely because it is consciously controlled. Often though, especially later in learning, the verbal system does not provide the optimal strategy to solve tasks that require the integration of information or cannot be described with a rule. In those cases, the procedural system will eventually outperform the explicit system and hence start to determine the categorical decisions. Initially, the explicit learning system was suggested to be closely linked to language, but because primates were shown to closely resemble humans in explicit learning behavior (see 1.1.2.3; Smith, 2010) this definition was adapted for work in animal models.

The authors also provide a suggestion of the individual algorithmic steps that each system employs. Hereby, the explicit system learns categories by identifying a number of possible rules, iteratively learning to select the optimal rule and learning a criterion value along this rule. In parallel to this explicit learning, the implicit system maps individual stimuli to category - and therefore motor - responses. Through convergence of information and repetition, a perceptual space slowly becomes associated with a category. Ashby et al. (1998) suggest that these two systems compete for the decision output through lateral inhibition, by which the system that is activated more strongly inhibits the other system.

Because COVIS was constructed based on neuropsychological and neuroimaging results, Ashby et al. (1998) give concrete predictions on which system involves which brain areas. The explicit system engages prefrontal and anterior cingulate cortices and part of the striatum, the head of the caudate nucleus. According to the prediction, the prefrontal and cingulate areas perform rule selection, consistent with findings that cingulate areas are involved in rule selection and Stroop tasks (Posner and Petersen, 1990; Bench et al., 1993). The criterion value along the selected rule is learned iteratively by the striatum. Initially, the COVIS model did not propose an involvement of the hippocampus or medial temporal lobe areas. However, increasing evidence from neuroimaging studies linked hippocampal activation to prefrontal learning of explicit categorization (Nomura et al., 2007).

The implicit system strongly relies on the tail of the caudate nucleus and its projections from sensory areas, such as ITC. Through repetition and dopamine-mediated reward learning

(Beninger, 1983; Wise and Rompre, 1989), the caudate nucleus maps ITC input onto a motor response. This process has been modeled by the decision bound theory (see 1.1.1.3; Maddox and Ashby, 1993) and in later studies as a striatal pattern classifier (Ashby et al., 2007). In both explicit and procedural systems, a part of the striatum will give the categorical decision as an output. These striatal outputs compete in their strength, reflecting the confidence of the respective categorical decision. Through lateral inhibition within the striatum, potentially enhanced by dopamine (Wickens et al., 1991), the stronger system will inhibit the other one and resolve the competition.

The COVIS model makes several behavioral predictions (for review see Maddox and Ashby, 2004). One prediction is that the absence or delay of feedback specifically impairs procedural learning because the proposed learning processes in the striatum rely on dopamine signaling of reward (Beninger, 1983; Ashby et al., 2007). This prediction was experimentally confirmed by comparing performance in rule-based and information-integration tasks without feedback (Ashby et al., 1999; Maddox et al., 2003).

COVIS further predicts that a change in motor requirement or visual-field position affects information-integration categorization performance more than rule-based performance, because procedural learning in the striatum closely links the input from visual cortical neurons to a motor output. Both suggested manipulations, changing the motor response (Ashby et al., 2003) and the retinotopic location of stimulus presentation (Rosedahl et al., 2018) did indeed specifically impair learning of implicit categorization.

On the other hand, since explicit category learning relies on prefrontal areas and attentional processes, interference with attention through another task should specifically affect rule-based category learning. Waldron and Ashby (2001), and later others (Zeithamova and Maddox, 2006; Filoteo et al., 2010) performed such dual task interference experiments and confirmed the effect. Participants were trained in rule-based or information-integration tasks, either as a single task or while the participants were performing another working memory intensive task, like the Stroop test, in parallel. Rule-based category learning under this condition was slowed significantly compared to the single task learning.

In summary, the COVIS theory of category learning describes an algorithmic and neuropsychological account of two parallel learning systems that compete in driving categorical decisions. It recapitulates behavioral and neuroimaging findings and gives specific testable predictions, like dopamine levels in striatum and the time course of the involvement of certain brain areas. However, it only provides limited suggestions on how neurons in these brain areas contribute to category learning. Several models have since been put forward, specifically focusing on predicting individual neuron activity and the interaction of neurons that could underlie the learning and computation of categories. I will consider examples of such models in the following section, specifically highlighting a recent model based on findings from electrophysiological recordings in primates.

1.2.4.3 Neuronal circuit models of category learning

One of the first neurocomputational models attempting to create an account of category learning on the level of individual neurons was proposed by Knoblich et al. (2002). This model was an extension to an object recognition model, called HMAX (Riesenhuber and Poggio, 1999), based on electrophysiological data from inferior temporal cortex (Logothetis et al., 1995). This model implemented a layer of neurons in the visual processing pathway, that performed a linear sum of input features and a following layer that performed a 'MAX' operation, i.e. each neuron being driven only by its strongest input. Subsequently another linear and another non-linear, 'MAX' layer were added. By combining the linear and nonlinear layers, the model created neurons that were tuned to visual objects invariant to position in visual field or scale, reproducing the observed tuning in ITC. Inputs from such object-tuned neurons were linearly combined and used to train 'categorization units' in a supervised fashion. These categorization units matched the recorded neuronal responses in prefrontal cortex of trained monkeys (Freedman et al., 2001). The authors conclude that the neuronal responses to category stimuli that can be observed in area ITC reflect the unsupervised learning of object representations based on the statistics of its inputs. In contrast, the category selectivity in PFC arises from supervised learning of a linear combination of object-tuned inputs and task-specific category information based on behavioral relevance. Pannunzi et al. (2012) extended this model by including an influence of behavioral relevance on learning in ITC by adding a top-down connection from PFC to ITC that modulated the responses in ITC based on the learned features for categorization (data from Sigala and Logothetis, 2002). Such a modulation was implemented with synaptic plasticity based on the correlation structure of pre- and postsynaptic activity, i.e. Hebbian and anti-Hebbian, upon correct and incorrect category decisions, respectively, and resulted in faster and more robust learning. However, whether such computations are indeed implemented by the connectivity between ITC and PFC has to date not been tested.

Neurocomputational implementations of COVIS (see 1.2.4.2) focused on modeling how two category learning systems, explicit and implicit, could learn potentially different categories and compete for the category decision. A model by Paul and Ashby (2013) assumes, based on experimental data from Ashby and Crossley (2010), that the control over the category decision does not switch between explicit and implicit categorization systems on a trial-by-trial basis. Rather, if the implicit category system performs better than the explicit one, the control over behavior switches only once towards the implicit system. As a consequence, the procedural learning system has to learn in parallel to the explicit system, although both systems likely receive the same feedback. In order for the procedural system to learn, even though the explicit system determines the behavior, the explicit system must inform or ‘teach’ the procedural system of the decisions and their outcome trial-by-trial. One prediction of the model is that neurons of the procedural learning system in the striatum receive an efference copy of the decision of the explicit system, either from prefrontal areas or another part of the striatum. This efference copy would drive the activity of striatal neurons within the procedural learning system whose response aligned with the decision of the explicit system. Hereby, their activity could be associated with a following dopamine signal, i.e. reward. Although Paul and Ashby (2013) hypothesize that the efference copy could reach striatum via premotor or motor areas, evidence for such a functional connection is yet to be found, and it is currently unclear how those inputs would target the relevant striatal units.

Another, more recent neurocomputational model focuses on interactions between prefrontal cortex and striatum during category learning. Building on the observed involvement of striatum in almost all categorization tasks and the finding of category-selective neurons in prefrontal cortex, Villagrasa et al. (2018) aimed to better understand the circuit mechanism of the interaction between these two areas (Fig. 1.3). Their model was informed by data from electrophysiological recordings in both areas (Antzoulatos and Miller, 2011) showing early category information in striatum that decreased as more stimuli were learned and category selectivity in prefrontal cortex that increased with training. The hypothesis is that striatum builds up first stimulus associations from its connections with ITC, and therefore forms a rudimentary category representation. Then, due to its connectivity with thalamus, the striatum ‘teaches’ prefrontal neurons by disinhibiting thalamus and thereby biasing prefrontal activity in order to facilitate Hebbian plasticity between ITC and PFC neurons. Indeed, early in training many modelled striatal neurons developed stimulus selectivity and fewer neurons showed category selectivity. With learning of more novel stimuli, the striatal variability in responses to a category increased, leading to a decrease in overall category selectivity. In PFC on the other hand, stimulus selectivity stayed low throughout learning, but category selectivity increased without the increased variability in response to novel stimuli. The resulting category representation in prefrontal cortex hence developed slower, but was generalized better to novel stimuli.

In summary, several neurocomputational models have been developed to fit a data set or experimental finding from human or non-human animal category learning research and to create testable predictions. However, often the models focus on neurons within a specific brain area or on the interaction between only two areas. So far, there has been no account of a unified picture of how the brain could implement category learning. One possible reason could be that the existing models are mainly based on human neuroimaging or primate electrophysiology data that offer limited insight into the underlying circuitry. Information about specific cell types, the functional connectivity between areas or the changes throughout the learning process is harder to obtain in these model systems. To date, there have been only few efforts to study category learning in mice and to connect results to human or primate models. However, mice offer better access to genetic and optical interrogation tools to test circuit predictions of category learning models and to thereby build a coherent picture of category learning in the brain.

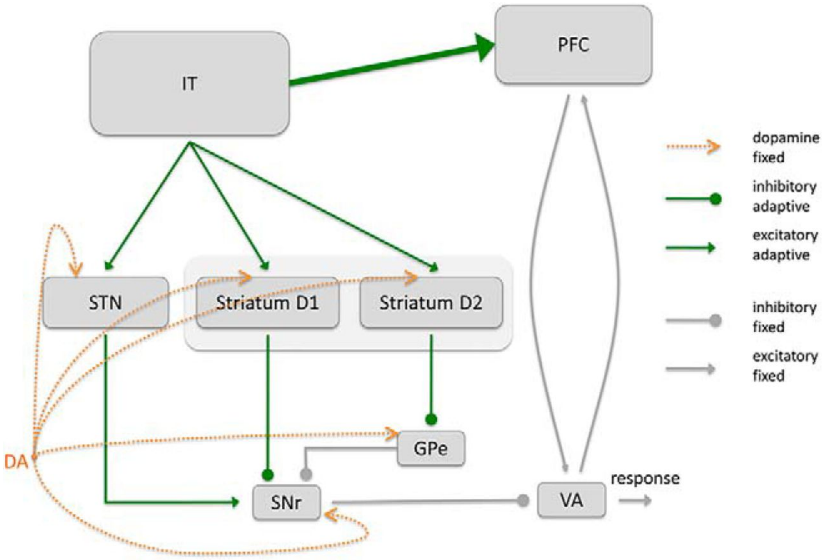


Figure 1.3: **Network components of striatal teaching model.** Through dopamine-mediated (orange) plasticity at the IT-striatum connections (green), the striatum will form a representation of learned stimuli early on. Fixed connectivity (grey) of the basal ganglia with the ventral anterior nucleus (VA) of the thalamus and of VA with PFC enables the basal ganglia to bias activity in PFC. In mice, the thalamic nucleus is termed ventral medial (VM) nucleus, but shows the same connectivity with the substantia nigra and prefrontal cortex (Kase et al., 2015; Collins et al., 2018). The resulting biased activity in PFC drives Hebbian (unsupervised) learning at the IT-PFC synapses and hence a later development of a category representation in PFC (Reprinted from Villagrasa et al., 2018).

1.3 A mouse model of category learning

1.3.1 Why use the mouse model?

As described in the previous sections, a substantial body of research has built up behavioral, algorithmic and also neurocomputational models of category learning. Many model predictions have been tested in humans with behavioral studies of healthy subjects and patients, and neuroimaging experiments. Further, single-unit recordings in primates provided the resolution to observe individual neurons, that human neuroscience investigation tools typically lacked. However, neuroscience research in primates is short of the versatility of observation and interrogation tools developed in rodent, especially mouse, neuroscience research.

The large-scale tool development in the mouse model was kickstarted with the discovery of techniques for the manipulation of genetic material and hence the creation of transgenic mice (Thomas and Capecchi, 1987). Due to their fast reproductive cycle and cheap maintenance, a high number of animals could be bred, screened and characterized. Such a method to generate knock-in or knock-out mouse lines of any desired gene brought control over specific molecular pathways, cell types and neuronal activity through chemical (Armbruster et al., 2007; Roth, 2016) or optical means (Boyden et al., 2005; Deisseroth, 2015). In addition, the fairly small brain of mice, compared to larger rodents, cats and primates, enabled the fast, large-scale observation and manipulation of neuronal activity *in vivo* with high resolution through techniques like two-photon microscopy (Denk et al., 1990, see 1.3.3). All those advantages have put the mouse into the center of biomedical and circuit neuroscience research. However, for the most part mice have not played a significant role in cognitive neuroscience. Only in the recent years, researchers have aimed to develop mouse models for more cognitive behaviors.

In category learning, a mouse model that can replicate the key findings from human and primate research could give three key benefits. First, it allows researchers to identify and chronically record specific neuronal projections through category learning, hence directly observe the interaction between several brain areas. Second, inputs to a brain structure can be anatomically traced in order to come up with better predictions of category learning networks. Third, the activity of neuronal populations or specific projections can be opto- or chemogenetically manipulated to test predictions of neurocomputational models. Thus in the neuroscience of category learning, the mouse model could be at a bridging point between behavior that is comparable to humans and a battery of neuroscience tools that can advance our understanding of category learning in the brain.

1.3.2 Designing a behavior paradigm for mice

1.3.2.1 Complex cognitive behaviors in mice?

Does it make sense to probe cognitive behaviors in mice? In order to answer this question one can consider the following two things: whether mice can learn behaviors comparable to humans, and whether there are any indications that the mechanisms underlying such behaviors can be compared. There are three hypothesized scenarios.

The first hypothesis is that mice cannot perform complex cognitive behaviors that can be compared to humans. From that, it follows that mouse behavior can be studied for its own merit, but not for learning about human behaviors and their underlying brain mechanisms. This view is supported by the observation that mice often require very long training times to learn tasks (Colacicco et al., 2002) or use of different strategies (Lipp and Wolfer, 1998) compared to rats. This lead to the majority of cognitive neuroscience studies being conducted in rats and non-human primates. Only in the recent years, more and more complex behaviors have been tested in mice, challenging the idea that mice do not show cognitive behaviors.

The second hypothesis is that mice can perform cognitive behaviors that are comparable to humans, but solve them with a different mechanism. Possibly, this hypothesis applies to categorization behavior in pigeons. Pigeons can categorize visual stimuli to very high accuracies, even natural images and human faces (Herrnstein and Loveland, 1964). However, when comparing humans and pigeons in implicit versus explicit categorization tasks, pigeons show different behaviors to humans, indicating that the two species use different mechanisms for category learning (see 1.1.2.2, Smith et al., 2012). This finding makes the pigeon model of category learning less comparable to humans. While studies across a wide variety of behaviors and brain structures are important for understanding specific model systems and learning how similar problems can be solved in different ways, it is debatable how much we can learn about mechanisms in the human brain.

The third hypothesis is that mice can perform cognitive behaviors that can be compared to humans and that the same brain circuits and mechanisms are involved. Behavioral and neuronal results that are comparable to findings from primate studies support this hypothesis, such as the observation of category representations in mouse PPC (Zhong et al., 2019) or the finding that mice can learn complex rule-switching tasks (Rikhye et al., 2018). However so far, category learning studies in mice have not aimed at using similar category structures, nor tried to compare neuronal findings to primate or human literature and models.

In conclusion, the question of whether mice show complex cognitive behaviors is not completely answered. However, prior studies of category- and rule-learning highlight the use of a mouse model for category learning that provides the comparison of key behavioral and neuronal results to primates and humans.

1.3.2.2 Optimizing behavioral studies in mice

What makes experiments in mice comparable to human experiments? And what makes a mouse experiment useful for gaining insights beyond results from human studies? A number of considerations about experimental design will influence the behavioral and neuronal results and therefore have a large impact on these questions (see also Discussion).

The first consideration is the motivational incentive in a learning task. Studying category learning involves engaging experimental subjects in an operant conditioning task (Skinner, 1935). The most frequent design of an operant learning paradigm across species is appetitive reinforcement learning. Hereby, a correctly displayed behavior will be reinforced with a reward, like a snack or money. In primate or rodent research, food, soymilk or water rewards are commonly used to reinforce the operant behavior of the animals (Guo et al., 2014). However, often the food or water reinforcer alone is not strong enough to motivate the animals to participate in the experiment (Dickinson and Balleine, 1994). Therefore, animals can be put on a restriction regime, limiting their food or fluid consumption, in order to make the food or water rewards a stronger motivational incentive (Tucci et al., 2006). Since both, the restriction regime as well as the choice of a reinforcer for the experiment, can affect the motivation of the animals (Berridge, 2004; Tucci et al., 2006) and may engage different brain circuits (Jourjine et al., 2016; Jourjine, 2017), they can influence behavioral and neuronal findings of a learning experiment and therefore potentially confound the results.

Another important aspect is the tested sensory modality. Most category learning studies in humans and primates are using visual stimuli. Our understanding of how we decompose visual scenes into individual features and how these features are separated and represented in the brain is better than for other sensory modalities. Hence, by testing category learning with visual stimuli, we have better control over individual stimulus features and thereby the type and difficulty of the categorization. However, when translating these experiments to rodents, it is important to keep in mind that mice have a far lower visual acuity (Sinex et al., 1979) and therefore might have difficulties perceiving stimuli that are used in human or primate category learning tasks.

Finally, the choice of operant task design will likely influence both the behavior (Guo et al., 2014) and the brain (David et al., 2012; Kuchibhotla et al., 2017). Operant behaviors that are closer to the natural repertoire of an animal will likely be easier to reinforce and therefore faster to learn, like in mice foraging behavior in freely-moving task conditions. An alternative to freely-moving paradigms is a task design that involves head fixation of the animal. Such tasks are frequently employed in primate and mouse neuroscience, because they give experimenters better control over the sensory stimulation and decrease the complexity of displayed behaviors. Hence, head fixation can reduce potential confounders to neural recordings and also enable calcium imaging with two-photon microscopes during the behavioral experiments. Operant behaviors in head-fixed tasks for mice can involve licking on a water spout (Guo et al., 2014), operating a choice ball (Sanders and Kepecs, 2012) or running on a treadmill (Hölscher et al., 2005). Even though there are efforts to develop more naturalistic head-fixed tasks (Havenith et al., 2018), most of these operant behaviors are not as natural and cause more stress to mice (Juczewski et al., 2020) than what is typically required of humans performing a task, e.g. pressing keys on a keyboard. This could affect the learning speed or even the learning strategy in a task and impact comparability.

In summary, developing a category learning task in the mouse requires choosing a motivational incentive, a sensory modality to test and an operant behavior. All of these decisions likely impact the behavioral and neuronal findings of the experiments and therefore also the comparability between results from humans, primates and mice. The first study I present in this thesis characterizes the effect of two training regimes with different motivational incentives on animal welfare and learning performance, establishing the optimal training parameters for category learning in mice.

1.3.3 Imaging neuronal populations in the mouse brain during behavior

Recording neural activity during behavioral testing was traditionally done in both primates and rodents using electrophysiology of single or multiple units at a time. While electrophysiological recordings still provide the highest temporal resolution, three major advances have made imaging of neuronal activity an indispensable recording technique, especially in mice.

First, fluorescent calcium indicators enable the optical detection of neuronal activity (Yuste and Denk, 1995; Stosiek et al., 2003). Calcium indicators either change their fluorescence intensity or absorption and emission spectra upon binding calcium ions and thereby signal calcium influx into the neuron, a proxy for action potentials. Further, GFP-derived calcium indicators, e.g. GCaMP, can be genetically encoded for long term expression in selected neuronal populations (Miyawaki et al., 1997). Such genetically encoded calcium indicators (GECIs) can be combined with structural markers, i.e. fluorescent molecules that do not vary with calcium concentration. Structural markers enable long-term observation and identification of neuronal populations even for individual neurons that do not show calcium activity at a given time (Mank et al., 2008; Tian et al., 2009; Chen et al., 2013; Rose et al., 2014).

Second, two-photon microscopy enabled the chronic recording of activity with subcellular resolution. Because it relies on the two-photon effect (Göppert-Mayer, 1931), light absorption and fluorescence are restricted to a narrow plane in the tissue. Therefore, two-photon microscopy causes less photobleaching and photodamage than one-photon microscopy with comparable resolution, and enables imaging deeper in brain tissue (Denk et al., 1990; Helmchen and Denk, 2005; Svoboda and Yasuda, 2006).

Third, a chronic glass implant can replace a certain area of skull above a brain region, or the entire dorsal surface of skull (Kim et al., 2016), allowing for recordings through that glass window for weeks to months (Holtmaat et al., 2009). However, the depth limit of two-photon microscopy restricts recordings with cranial window implants to the dorsal (~500µm) surface of the brain. Thus, any brain area that is not on the surface, including areas of the medial prefrontal cortex, cannot easily be reached and imaged.

One common technique to overcome this limitation is the implantation of a cannula or a gradient refractive index (GRIN) lens above the area of interest. With such an implant, areas like the hippocampus or the striatum and also medial prefrontal cortex have been made accessible for two-photon microscopy (Barretto et al., 2009; Pinto and Dan, 2015). However, such an implant either displaces the tissue above the area of interest, which partially lesions surrounding tissue, or requires aspiration of the tissue, creating a major lesion to any tissue above the area of interest. Another, less common technique is the implantation of a mirror-coated microprism that can be fit into fissures of the brain, e.g. between cerebrum and cerebellum to observe entorhinal cortical areas or between the hemispheres of the frontal lobes to observe medial prefrontal cortex (Low et al., 2014). With this technique, tissue surrounding the area of interest is not lesioned by the glass prism, but rather displaced, with no worse effect on the brain than a cranial window implant.

In summary, while electrophysiological recordings provide superior temporal resolution, two-photon microscopy in combination with GECIs and chronic window or microprism implants enables the observation of activity of individual neurons over months in awake, behaving animals. Therefore, chronic two-photon calcium imaging in mice allows for investigating changes in neuronal activity in medial prefrontal cortex throughout a behavioral learning period, like category learning.

1.3.4 Investigating the mouse brain during category learning

The two previous sections consider more technical aspects of operant conditioning (see 1.3.2.2) and simultaneous recording of neuronal activity (see 1.3.3), aiming to utilize the toolkit of mouse neuroscience research while maintaining comparability to studies in humans and non-human primates.

In this last section, I want to put these considerations into the context of category learning research in the mouse and, thereby, highlight the aims of the studies presented in this thesis.

1.3.4.1 The role of sensory and parietal areas in mouse categorization

So far, studies that make use of the mouse model by combining category learning with neuronal recordings and manipulations exclusively involve auditory categorization tasks. In these studies, mice were tasked to discriminate high frequency tones from low frequency tones. The boundary between the two categories could either lie in the middle between the presented frequencies, or shift closer to higher or lower tones, enabling testing of flexible categorization. Neuronal recordings were performed in auditory cortex (Xin et al., 2019) and posterior parietal cortex (Zhong et al., 2019). Auditory categorization modulated the responses to the stimuli in auditory cortex and PPC and activity in PPC was crucial during, but not after, category learning. These auditory categorization tasks presented only one dimensional stimuli (frequency of tones), so these tasks would be considered rule-based category learning tasks, but could not be used to investigate selective attention, a key component of explicit categorization (see 1.1.2.2). Such a design also made it difficult to integrate the obtained neural results into the existing models of category learning from humans and primates.

Visual categorization in mice has so far only been tested in behavioral studies without neuronal recording. Watanabe (2013, 2017) has found that mice are able to learn categorizing images of paintings based on the painter, i.e. identifying Kandinsky's paintings from Mondrian's. In another study, Creighton et al. (2019) trained mice in an object category recognition task, and found that mice could distinguish between novel objects that were part of a formerly experienced object category, e.g. 'car', and objects that were of an unfamiliar category. Both of these experiments studied visual category learning in mice, but were restricted to high dimensional stimuli (paintings and 3D objects) and thereby tests for implicit categorization. Explicit visual categorization, with precise control over the stimulus dimensionality, is yet to be tested in mice and will provide an important piece of the puzzle to compare category learning in mice to the existing body of research in rats, pigeons, primates and humans.

1.3.4.2 Investigating mouse prefrontal cortex in category learning

In the mouse, the prefrontal cortex has not been explored with respect to category learning. However, it is known to encode task-relevant cues, like stimuli or reward presence (Pinto and Dan, 2015) and learned rules (Rikhye et al., 2018). Considering those findings and the specific involvement of PFC in rule-based category learning in rats (Broschard et al., 2021), it is likely that also in mice prefrontal cortex plays an important role in explicit category learning. Possibly, in trained animals prefrontal cortex holds a representation of the learned categories. A neural representation of categories or learned rules for categorization could, on the one hand, involve individual neurons showing category-selective activity, similar to the findings in non-human primates (Freedman et al., 2001) or the type of coding observed in mice after learning a task rule (Rikhye et al., 2018). On the other hand, categories could be rather represented with a type of distributed code across a neuronal population, as supported by the observed mixed selectivity for task-relevant parameters (Rigotti et al., 2013).

Although there is still an ongoing debate about the anatomical relations between human, primate and rodent prefrontal cortical areas, the many functional analogies between primate and mouse prefrontal cortex (for review see Carlén, 2017; Merre et al., 2021) suggest that findings from mouse category learning could be generalized to the human. Since there are promising theories and models built from human and primate category learning research, and the mouse model offers a much broader variety of investigation tools, investigating mouse prefrontal cortex during category learning could help us improve the understanding of human category learning and underlying brain mechanisms.

Therefore, the focus of the second study presented in this thesis is to establish a category learning paradigm in the mouse that enables comparability to existing findings from primates and involves simultaneous neuronal recording of PFC populations. In order to go beyond the current understanding of the role of PFC in category learning, I follow individual neurons in mouse PFC throughout the entire learning paradigm of the animals using chronic two-photon calcium imaging during the task. From these data, I aim to identify neuronal representations of categories and other task-relevant parameters as well as characterize their emergence with learning.

2 | Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice

RESEARCH ARTICLE

Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice

Pieter M. Goltstein[1☯*], **Sandra Reinert**[1,2☯], **Annet Glas**[1,2☯], **Tobias Bonhoeffer**[1], **Mark Hübener**[1*]

1 Max Planck Institute of Neurobiology, Martinsried, Germany, **2** Graduate School of Systemic Neurosciences, Martinsried, Germany

☯ These authors contributed equally to this work.
* goltstein@neuro.mpg.de (PG); mark@neuro.mpg.de (MH)

Abstract

Head-fixed behavioral tasks can provide important insights into cognitive processes in rodents. Despite the widespread use of this experimental approach, there is only limited knowledge of how differences in task parameters, such as motivational incentives, affect overall task performance. Here, we provide a detailed methodological description of the setup and procedures for training mice efficiently on a two-choice lick left/lick right visual discrimination task. We characterize the effects of two distinct restriction regimens, i.e. food and water restriction, on animal wellbeing, activity patterns, task acquisition, and performance. While we observed reduced behavioral activity during the period of food and water restriction, the average animal discomfort scores remained in the 'sub-threshold' and 'mild' categories throughout the experiment, irrespective of the restriction regimen. We found that the type of restriction significantly influenced specific aspects of task acquisition and engagement, i.e. the number of sessions until the learning criterion was reached and the number of trials performed per session, but it did not affect maximum learning curve performance. These results indicate that the choice of restriction paradigm does not strongly affect animal wellbeing, but it can have a significant effect on how mice perform in a task.

Citation: Goltstein PM, Reinert S, Glas A, Bonhoeffer T, Hübener M (2018) Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice. PLoS ONE 13(9): e0204066. https://doi.org/10.1371/journal.pone.0204066

Editor: Manabu Sakakibara, Tokai University, JAPAN

Received: July 24, 2018

Accepted: August 31, 2018

Published: September 13, 2018

Data Availability Statement: The data are available on https://web.gin.g-node.org/pgoltstein/HeadfixedLearning_Goltstein_Huebener_2018.

Funding: This project has received funding from the Max Planck Society and the Collaborative Research Center SFB870 of the German Research Foundation (DFG) to T.B. and M.H.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Rodents, in particular rats and mice, have long been used in behavioral studies exploring the mechanisms underlying learning and memory [1,2]. Such experiments are particularly valuable when combined with simultaneous recordings from neurons involved in the task. Traditionally, this is done with extracellular recordings of single- or multi-unit activity, a technique that can easily be adopted to freely moving animals [3]. In some instances, however, it is desirable to carry out behavioral experiments in movement-restricted animals. Head-fixation in particular is indispensable under certain conditions, e.g. when precise control over sensory inputs is needed, or when the employed recording technique is sensitive to brain motion, like patch clamp recordings [4] and two-photon microscopy [5].

Head-fixed operant conditioning is now commonly used to train mice in diverse sensory detection- and discrimination tasks, as well as in virtual navigation experiments. Such tasks can be performed using various operant-stimulus modalities, e.g. visual [6], auditory [7], olfactory [8] or tactile [9]. The most widely used paradigm is the Go/No-Go task, in which the animal makes a choice by either performing or withholding from a certain behavior, such as a lick on a lick spout [9], a lever press [6] or running [10]. An important factor in behavioral training, especially with parallel neuronal recordings, is the ability to differentiate between the actual choice of a mouse and the mere level of motivation to participate in a task. Go/No-Go task designs lack the ability to precisely differentiate between an active No-Go (active withholding) and a passive No-Go, reflecting loss of motivation. Two-choice designs are therefore often more appropriate as they better allow discriminating between active choices of a mouse, e.g. licks left or right [9] or steering wheel movements to the left or right [11], and its task engagement (finished versus missed trials). Head-fixed paradigms also vary in the dimensionality of body movement that is permitted and measured. While some virtual reality approaches allow more degrees of freedom [12–14], it is common to restrict running to one dimension [10,15] or restrict body movement entirely by placing the animal in a narrow tube [9,16]. Beyond these, many more detailed parameters, e.g. setup design, training protocol, trial sequence and stimulus presentation can be adjusted to suit the specific experimental need.

The effect of such parameter choices on the outcome of a behavioral experiment is often not systematically explored and only occasionally reported in the literature. One such parameter is the choice of (naturalistic) motivational incentive. This can be appetitive (e.g. reward) or aversive (e.g. fear) and is commonly administered by delivery of food or water [9] or by delivery of mild shocks, respectively [17]. Animal behavior can also be motivated using targeted, optogenetic activation of dopaminergic circuits [11] or by circuits driving hunger or thirst [18]. Still, head-fixed learning paradigms mostly use food and water restriction, in part because it does not require additional optical equipment. While food and water restriction regimens are sometimes perceived as interchangeable, these two methods engage the animal's physiology differently [19,20], and hunger and thirst recruit different neuronal circuits [18,21,22]. Therefore, similar levels of food and water restriction, as usually measured by the animal's relative reduction in body weight, might affect task performance, task motivation and also animal welfare in a different manner [23].

This study provides a detailed description of the setup design and procedures to efficiently train mice using either food or water restriction on an appetitive operant visual discrimination task. We explicitly monitor animal welfare using measurements of body weight and a standardized scoring routine, as well as continuously recorded physical activity patterns from the home cage [24]. We demonstrate the sensitivity and reliability of our conditioning method by addressing how the choice for food or water restriction affects performance in head-fixed operant conditioning.

Methods

Animals

All procedures were performed in accordance with the institutional guidelines of the Max Planck Society and the local government (protocol number 55.2-1-54-2531-213-2015, approved by the Beratende Ethikkommission nach § 15 Tierschutzgesetz, Regierung von Oberbayern). Twelve male C57BL/6J mice (postnatal day 34) were individually housed in standard individually ventilated cages (IVC; Tecniplast GM500) and placed in a Digital Ventilated Rack (DVC, Tecniplast). Each cage was equipped with a dedicated electronic board (DVC board) composed of 12 electromagnetic field generating electrodes evenly positioned in a 4 by

3 grid underneath the entire cage floor area. Sensors measured activity at each electrode separately (4 Hz sampling frequency) and stored the data on a computer. Disturbances in the strength of the local electromagnetic field were used as proxy for a mouse's behavioral activity in the home cage (see Data analysis). All mice were kept on an inverted 12-h light, 12-h dark cycle with lights on at 22:00. Ambient temperature (21.0 ± 0.7 °C) and humidity (63 ± 2%) were kept constant. Water and standard chow (Altromin Spezialfutter GmbH, #1310) were provided to the mice *ad libitum* prior to behavioral experiments. Starting seven days before surgery, mice were handled and weighed daily by the same experimenters (two female, two male) that later also carried out behavioral training. After completion of behavioral procedures, mice were euthanized using CO_2 asphyxiation.

Surgical procedures

Mice were anesthetized with a mixture of fentanyl, midazolam and meditomidine in saline (0.05mg/kg, 5 mg/kg and 0.5 mg/kg respectively, injected i.p.) and sufficient depth of anesthesia was confirmed by absence of the pedal reflex. Eyes were covered with a thin layer of ophthalmic ointment (IsoptoMax). Lidocaine (0.2mg/ml) was applied onto the scalp for topical anesthesia and carprofen in saline (5mg/kg, injected s.c.) was administered for analgesia. The skull was exposed, dried and scraped with a scalpel to facilitate attachment of the head plate. The custom-designed head plate (Fig 1D; S1 File) was fixed in position, over the left parietal bone, using cyanoacrylate glue and subsequently secured with dental acrylic (Paladur). After surgery, mice were injected with a mixture of the antagonists naloxone, flumazenil and atipamezole in saline (1.2 mg/kg, 0.5 mg/kg and 2.5mg/kg respectively, injected s.c.) and left to recover under a heat lamp. For post-operative analgesia, mice received carprofen (5mg/kg, injected s.c.) for three subsequent days.

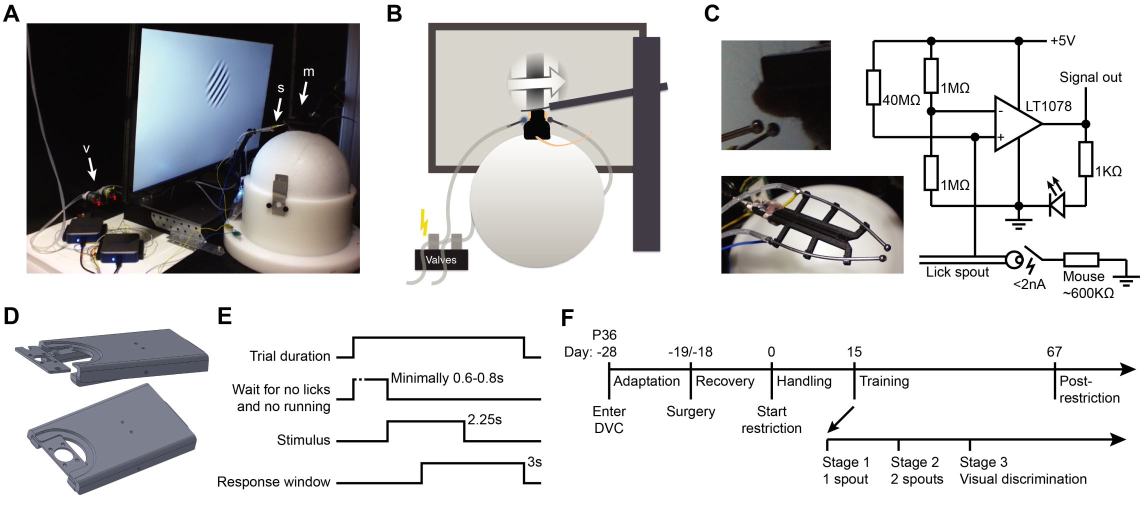


Fig 1. Behavioral apparatus and training protocol. A. Setup used for head-fixed visual conditioning. Arrow 'm' points to a head-fixed mouse, resting on a Styrofoam ball, in front of a centrally positioned monitor and the two lick spouts (arrow 's'). Arrow 'v' indicates the pinch valves for reward delivery. B. Schematic of the behavioral setup as seen from behind. C. Lick detection. Top left: Position of dual lick spouts in front of the mouse. Bottom left: Photo of fully assembled 3D printed lick spout holder. Right: Electrical circuit for contact/lick detection on a single lick spout. D. 3D renderings of head-bar and head-bar holder. E. Temporal sequence of within-trial phases. Reward is delivered immediately upon the first (correct) lick in the response window. F. Overall experimental timeline depicting main experimental and training stages.

https://doi.org/10.1371/journal.pone.0204066.g001

Food and water restriction

Mice were randomly assigned to either the food-restricted or the water-restricted group. The period of restricted access to food or water was started 18–19 days after surgery. Animals were transferred to novel cages immediately before food or water restriction started.

At the start of water access restriction, mice were initially provided with 50% of the average *ad libitum* water intake per day (50% was on average 1.57 ml in our facility). The water ration was provided in the home cage using the nozzle of a standard water bottle that was closed off at the back using red tape. From the fourth day onwards, the water ration was first offered in a hand-held syringe during handling, with any remaining volume supplied in the home cage. In parallel, when a mouse reached the target weight of 85% of the initially measured *ad libitum* weight (reference weight), the daily volume of water supplement was individually adjusted in order to maintain the target weight. As precaution, a minimum daily ration was set to 25% of the *ad libitum* intake. However, the daily supplemented amount was only rarely as low as 25% of *ad libitum* intake. Water-restricted animals had *ad libitum* access to food throughout the experiment.

Food access was restricted according to the following procedure. On the first day, mice received the minimum ration size of 2.0 g standard chow (3.279 kcal/g) in their home cage. Subsequently, the daily ration was adjusted per mouse in order to keep its weight at 85% of the reference weight, while staying above the minimum ration weight of 2.0 g. From day four untill day seven, mice were fed unflavored soymilk (Alpro) from a handheld syringe during handling. However, we noted that mice were not particularly motivated to drink regular soymilk (average consumed volume per mouse; day 5: 0.58 ml; day 6: 0.63 ml; day 7: 0.51 ml). Thus, on days eight and nine we offered sweetened soymilk (Alpro), which did not increase the consumed volume (average consumed volume per mouse; day 8: 0.51 ml; day 9: 0.38 ml). Finally, from the tenth day until the end of the experiment we used infant formula soymilk (SMA Wysoy)[10]. The infant formula soymilk was prepared by adding lukewarm water to a falcon tube containing 10–12 ml of soymilk powder until the total volume of the suspension reached 30–34 ml. We immediately noticed a difference in consumption behavior when providing mice with infant formula soymilk (average consumed volume per mouse; day 7: 0.73 ml; day 8: 1.29 ml; day 9: 0.99 ml; day 10: 1.73 ml; day 11: 1.71 ml). The daily food ration of each mouse was reduced by an amount that matched the caloric content of the consumed volume of soymilk (0.67 kcal/ml). Food-restricted animals had *ad libitum* access to water during the entire experiment.

Animal welfare assessment

Daily welfare assessment involved scoring mice on five different aspects of wellbeing using individual scoresheets [9]. Scores on ‘Activity and behavior’ ranked an animal’s behavior in the home cage from normal, active (0), via reduced activity (1), only moves when touched (2) to lethargy (3). ‘Look/posture’ indicated the condition of the fur and the posture of the mouse, ranging from normal (0) to arched back and very shaggy fur (3). ‘Urine/feces’ was scored as indication of eating, drinking and associated physiological processes, ranging from normal (0), via reduced amounts (1) to none (2). ‘Body condition’ indicated the shape and outline of the mouse’s body and spine according to Ullman-Culleré & Foltz [25], ranging from normal (0), via underweight (1) to emaciated (2). ‘Signs of dehydration’ were assessed using skin turgor, ranging from none (0), via light (1), moderate (2), to strong (3). Finally, the cumulative score across all aspects was used to judge the overall wellbeing of the animal into four discomfort categories named according to the European Union based legislation. A discomfort score of zero was interpreted as ‘sub-threshold discomfort’, one as ‘mild discomfort’, between two and four as ‘moderate discomfort’ and higher than four as ‘severe discomfort’.

Apparatus for visual discrimination learning

Visual discrimination learning was carried out in custom-built setups that were placed in 75 x 75 x 75 cm boxes, providing a semi-enclosed environment (Fig 1A and 1B). The apparatus consisted of a head-plate holder, a spherical treadmill, a computer monitor, two lick spouts (16 Gauge, 3mm tip-diameter reusable feeding needles, Fine Science Tools) with lick detectors, tubes and valves to supply liquid reward. The treadmill was made of an airflow-supported Styrofoam ball [26] and restricted to forward and backward motion by a pin pushed into the side of the ball. An optical sensor, extracted from a computer mouse (G502, Logitech), tracked rotation of the ball using a custom-written LabVIEW (National Instruments) program. The mouse was head-fixed on the ball using a surgically implanted aluminum head plate, clamped into a custom-designed holder (S1 File). The head plate holder employed a (simplified) system of kinematic mounts to ensure reproducible positioning of the animal's head within the apparatus (Fig 1D; S1 File; [27]). Visual stimuli were presented on a gamma corrected computer monitor (Dell P2414H; resolution: 1920 by 1080 pixels; width: 52.8 cm; height 29.6 cm; maximum luminance: 182.3 Cd/m^2). The monitor was positioned in front of the mouse at a distance of 18 cm and centered at 0 degrees azimuth and elevation. The box was illuminated by red LEDs (630 nm), and a webcam (Logitech F100) was used to observe the mouse and setup within the enclosed space.

The two steel lick spouts were mounted on a custom 3D-printed holder that allowed fine adjustment of the space between the lick spout nozzles (S2 File). The lick spouts were positioned in front of the animals' mouth using a movable arm (Fig 1C, left panels). Care was taken to place the lick spout well within the reach of the tongue, which is especially important in the first pre-training sessions. Precise central positioning of the lick spouts with respect to the animal's mouth was critical; asymmetrical placement sometimes biased mice to make more licks on the closer spout. Each lick spout was connected to a custom-made lick detection circuit based on Weijnen [28] and Slotnick [29]. The circuit registered a voltage drop on the non-inverting high impedance input of an operational amplifier (LT1079CN; Linear Technologies) when the mouse short-circuited the input by licking on the spout (Fig 1C, right panel). The inverting input was connected to a voltage divider such that an individual lick triggered a strong discrete voltage drop in the amplifier output. The non-inverting and inverting inputs of the circuit could be switched in order for the circuit to report licks by voltage peaks. However, the described arrangement allows detecting whether the circuit is switched on from the baseline circuit output voltage.

Liquid reward was supplied through the lick spout by gravitational flow, operated using full opening pinch valves (NResearch Inc.). Valves were individually calibrated to supply drops of approximately 8 μl, which required valve-open durations of roughly 50 ms for water and approximately 75 ms for soymilk. Tubing was pressure-flushed with distilled water after each behavioral training session to prevent clogging. Signals from the lick detectors, the optical speed sensor and other triggers were recorded with two USB multifunction input/output devices (USB6001, National Instruments). The first device was used for closed loop control of the setup using a custom-written behavioral-training program (Matlab, Mathworks). The second device passively recorded all sensor signals at 500 Hz using a custom-written data-acquisition program (LabVIEW, National Instruments), which allowed for more precise offline analysis of behavioral parameters (see Data analysis).

Habituation and pre-training for head-fixed two-choice operant conditioning

Behavioral procedures were carried out six times per week between 14:00 and 18:00. In a two-week period prior to head-fixed operant training, mice were habituated to the experimental procedures. Each habituation session lasted 10 to 15 minutes during which the mouse was (1) held in the experimenter's gloved hands, (2) placed on the surface of a Styrofoam ball, (3) fed water or soymilk through a syringe and (4) accustomed to brief head fixation by holding the head plate manually for a few seconds. In this specific experiment animals were habituated for a period of two weeks because we tested different variants of soymilk (see above). However, mice typically accustom to these procedures in three to four days.

In order to shape animals for the head-fixed visual two-alternative choice task, we implemented two stages of head-fixed pre-training. The first stage familiarized animals with the association between timed licks and liquid reward from a single lick spout. To this end, animals were exposed to the trial sequence (Fig 1E), but in absence of visual stimulus presentation. Each trial started with an inter-trial interval of 2.0 s, after which the mouse was required to withhold licking and cease running (velocity below 1 cm/s) for a duration of at least 0.6 s to 0.8 s (varied per trial in order to prevent mice from learning a fixed timing sequence). When this requirement was met, the trial proceeded with the visual stimulus period. In pre-training stage 1 and 2 no actual visual stimulus was presented in this period, the screen remained blank. After 1.0 s from the onset of the visual stimulus period, the mouse could make a lick on the fluid spout in order to receive a single drop (approximately 8 μl) of water or soymilk. This period, during which a lick on the spout initiated reward delivery (named 'response window'), lasted initially 15.0 s and was gradually reduced to 5.0 s in subsequent pre-training sessions. At the start of the training sessions, a few drops were given by manual activation of the valves in order to motivate the mouse to lick for reward and to adjust the lick spout's positioning relative to the mouth and tongue. Mice proceeded to the second pre-training stage when they performed about 50 rewarded trials per training session on two consecutive days.

In pre-training stage 2 the trial sequence remained the same, except that now two lick spouts were positioned in front of the animal. On each trial, only a single lick spout was selected as active, and only a lick on this spout, during the response window, triggered reward delivery. Licks on the non-active spout were recorded but did not abort the remaining period of the trial/response window. The distance between the left and right lick spout was initially set to 1.0 to 1.5 mm. Later-on in pre-training stage 2, the inter-spout distance was increased to approximately 3.0–4.0 mm, the inter-trial interval was increased to 4.0 s and the response window duration was reduced to 4.0 s. Mice proceeded to the visual discrimination task when, in pre-training stage 2, animals performed a minimum of 50 trials per session and consumed drops without a strong preference for one of the two lick spouts.

Side bias correction strategy

Mice tend to develop a strategy of responding with a majority of the licks on only one of the two lick spouts (i.e. they showed a 'side-bias'), which we aimed to prevent using the following strategy. On each trial, we drew a random number r between -1 and +1. If this number was above an adjustable threshold t_b (bias-threshold), the next trial would give reward on the left spout, otherwise it would give reward on the right lick spout (Eq 1).

$$Next\ trial, side = (r > t_b \rightarrow Left) \wedge (r \leq t_b \rightarrow Right) \qquad \text{Eq 1}$$

The value of the threshold t_b was calculated using the outcome of the last 20 non-missed trials where $n_{correct\ left}$ and $n_{correct\ right}$ were the total number of trials in which the first lick in the response window was on the correct spout, and $n_{total\ left}$ and $n_{total\ right}$ were the total number of presented ‘left trials’ and ‘right trials’ within the 20-trial period (Eq 2).

$$t_b = \min\left\{m, \max\left\{-m, \left(\left(\frac{n_{correct\ left}}{n_{total\ left}}\right) - \left(\frac{c_{correct\ right}}{n_{total\ right}}\right)\right)\right\}\right\} \qquad \text{Eq 2}$$

The value of *m* instated a minimum probability for either stimulus to be selected by bounding the value of t_b to the range -*m* to +*m*. Thus, when a mouse would only lick on the left lick spout, the value of t_b would approximate *m*, reducing the chance that the next trial would be a ‘left trial’ to minimally $0.5 - {}^{m}/_{2}$ and increasing the chance that the next trial would be a ‘right trial’ to maximally $0.5 + {}^{m}/_{2}$. As a result, mice were presented with more trials on the non-preferred lick spout (i.e. right), gradually and eventually altering the mouse’s preference until it was balanced between spouts. The side-bias correction algorithm was active in pre-training stage 2 and during the first 5 to 7 training sessions of the visual discrimination stage.

Head-fixed visual two-choice operant conditioning

Two choice (lick left/lick right) operant conditioning featured visual stimuli consisting of sinusoidal gratings, drifting at 1.5 cycles/s. For each mouse, one stimulus was assigned to indicate ‘lick left’ and another to indicate ‘lick right’. These two stimuli were chosen such that each had one of two orientations that differed by 90 degrees, and each had either a low or a high spatial frequency (0.04 or 0.1221 cycles/degree). Selection of stimulus orientation and spatial frequency was counterbalanced across animals. Full contrast stimuli were presented in a 37 degree diameter circular area, centered at 10 degrees elevation and 0 degrees azimuth and blended within an annulus of 4 degrees width into an equiluminant grey background (total stimulus diameter including blended surround was 45 degrees).

Visual discrimination training followed the same basic trial structure as described above (Fig 1E), with the main addition that now a visual stimulus was presented for 2.25 s. The response window (3.0 s duration) started 1.0 s after stimulus onset. During the response window, a lick on the correct spout triggered reward delivery, while a lick on the incorrect spout caused a time-out. On rewarded/correct trials, stimulus presentation was continued for the full 2.25 s. On incorrect trials, the stimulus was replaced by a narrow horizontal black bar spanning the width of the display, presented for the duration of the time-out (2.5 s). Stimulus presentation or time-out was followed by an inter-trial interval of 5.0 s. Licks during the inter-trial interval and in the 1.0 s period between stimulus onset and response window onset (called ‘grace period’) [9] did not change the trial flow. In order to facilitate exploration and motivation, time-outs were not implemented in the first three training sessions. Therefore, in these initial sessions, an incorrect lick did not abort the response window and the mouse could still obtain a reward by subsequently licking on the correct spout.

Data analysis

Experimental and behavioral parameters such as the timing of licks, timing of drops, running speed, stimulus onset and other triggers were extracted from the passive data-recorder at 2 ms temporal resolution (LabVIEW, National Instruments) and analyzed using custom-written Matlab (Mathworks) and Python routines.

Continuous home cage activity patterns were calculated from the 12 sensors of the DVC system using a custom analysis program (written in Python). Each sensor provided a constant signal (4 Hz), which dropped when a mouse moved near/over it. The variance of the sensor signal was calculated within time bins of 1 minute and subsequently averaged across sensors, resulting in a minute-by-minute indication of average home cage activity per single housed animal. Per mouse/cage, outlying values (>95 percentile) were clipped to the value of the 95th percentile; these outliers often coincided with cage removal from or insertion into the rack. Next, all data points were normalized per mouse/cage (by division) to the 85th percentile of all values recorded during baseline periods (the 7 day period before surgery and the 14 day period before food/water restriction).

Learning performance was calculated as fraction of correct trials per session and evaluated across training sessions. The resulting learning curve was fit with a sigmoidal curve (Eq 3) where x was the average fraction correct trials per session, and parameters y_0 (minimum of curve), c (maximum of curve relative to y_0), k (steepness) and x_0 (time point of maximum steepness) were estimated using least squares fitting.

$$Fitted\ curve = y_0 + \frac{c}{1 + e^{-k(x-x_0)}} \qquad \text{Eq 3}$$

Latency to learning was determined as the number of sessions until an animal exceeded the criterion of 66% correct trials. The behavioral threshold of 66% correct trials was determined based on prior experience. The probability of detecting a single false-positive behavioral threshold crossing across the 23 sessions of the learning curve was 0.001 (assuming 100 trials per session). Maximum learning curve performance was estimated per mouse from the maximum of the individually fitted learning curve. This measure approximates the average level of performance that mice reached after 23 training sessions, independent of the latency to criterion. The total amount of water or soymilk that a mouse consumed during the task was computed from the number of drops that the mouse received. Data are presented as mean ±SD unless mentioned otherwise. Between-group statistical comparisons were carried out using a Mann-Whitney U test.

Results

Adaptation to reversed day/night cycle, surgery and recovery

Four weeks before starting food or water restriction, C57Bl/6j mice were transferred from a local animal breeding facility into individual 24hr/day activity monitoring cages (DVC, see Methods) that were kept in an animal holding room with a reversed day/night cycle. After an adaptation period of 9 to 10 days, the 12 animals with the highest bodyweight were randomly assigned to two experimental groups (food or water restriction, n = 6 each), implanted with a head-bar and subsequently allowed to recover until the start of the experiment. The four remaining mice (having the lowest body weight on the two surgery days) were not implanted and were left kept in their home cage throughout the experiment. While all implanted animals showed a reduction in body weight on the days immediately following surgery, both experimental groups recovered within seven days to a body weight that was comparable to the non-implanted group (Fig 2A).

Animal wellbeing during food or water restriction

The *ad libitum* reference weight of all mice was taken at 14:00 on day zero, after which food or water restriction was started (see Methods). The body weight of each mouse was maintained at around 85% of the individual *ad libitum* reference in all mice throughout the period of restricted food or water access (Fig 2A and 2B). All animals received a daily individually calibrated supplement of solid food (chow) or water (see Methods; Fig 2C) in addition to any soymilk or water they obtained during handling or training (Fig 2D).

Daily discomfort scores were assessed from the day of the surgery until 10 days after the end of food/water restriction (Fig 2E, solid lines). In addition, the institutes animal welfare

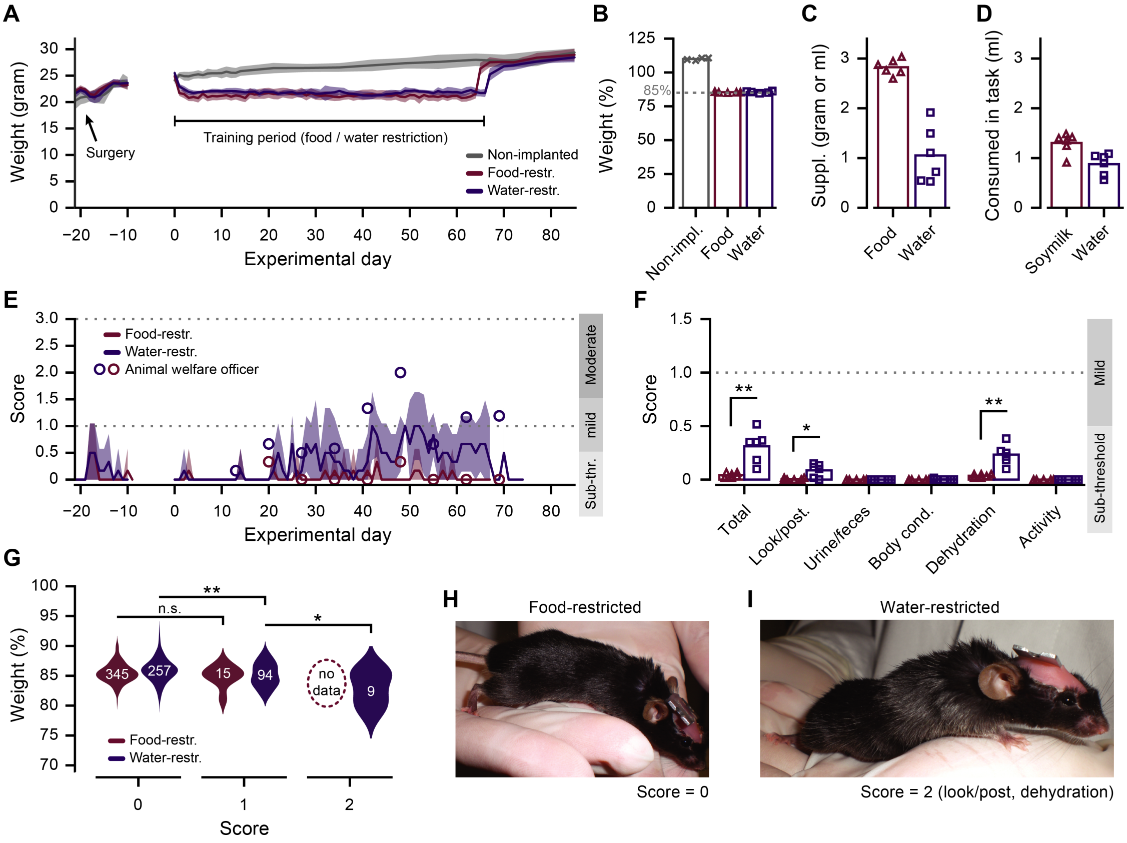


Fig 2. Animal weights and discomfort scores. A. Mean (±SD) daily weight of each experimental group across the entire experiment (gray: non-implanted, n = 4; red: food-restricted, n = 6; blue: water-restricted, n = 6). B. Average weight, in percentage of reference weight, throughout the period of food or water restriction. C. Amount of supplemented food (red) or water (blue) given (average of entire training period). D. Amount of soymilk (red) or water (blue) earned during training (average of entire training period). E. Mean (±SD) daily score of food (red) and water (blue) restricted mice over the entire experiment. Circles indicate scores as judged by the animal welfare officer. F. Daily score averaged across the period of food/water restriction. Total score is the sum across all five individual scores (look/posture, urine/feces, body condition, dehydration signs, activity; MWU test, *p = 0.018, ** p = 0.002). G. Distribution of daily measured weight as a function of the daily determined discomfort score for food- and water-restricted mice (MWU test, *p = 0.016, ** p = 0.002). Numbers in distribution plots indicate n in individual daily measurements. H. Example photo of food-restricted mouse (discomfort score 0, 'sub-threshold'). I. Example photo of water-restricted mouse (discomfort score, total = 2, 'moderate'; look/posture = 1; dehydration signs = 1). All panels: grey crosses (non-implanted), red triangles (food-restricted) and blue squares (water-restricted) indicate averages for individual animals.

https://doi.org/10.1371/journal.pone.0204066.g002

officer assessed scores weekly, each Monday at 09:00, during the period of food/water restriction (Fig 2E, circles). The daily assessment of discomfort during the course of the experiment ranged mostly from 'sub-threshold' to 'mild', but individual scores very occasionally exceeded into the 'moderate' range. Signs of discomfort were most often observed in the post-surgery period and from the third week of food/water restriction onwards (Fig 2E).

The average score during the period of food/water restriction remained well below the cut-off for 'moderate' discomfort (Fig 2F). However, the food-restricted group had significantly lower total scores compared to the water-restricted group (total: food-restricted, score = 0.04 ±0.02; water-restricted, score = 0.31±0.13; MWU test, p = 0.002; n = 12 mice). This difference was mostly caused by observations of mild skin turgor (signs of dehydration: food-restricted, score = 0.04±0.01; water-restricted, score = 0.23±0.09; MWU test, p = 0.002; n = 12 mice) and slightly erected, shaggy fur (look/posture: food-restricted, score = 0.01±0.01; water-restricted, score = 0.09±0.06; MWU test, p = 0.018; n = 12 mice). Scores on the other aspects did not exceed zero, except for a single occurrence of a score for the body condition of a water-restricted mouse.

The body weight of water-restricted mice with a total score above zero was on average significantly lower than that of mice with a zero score (score 0: 85.9±1.9%; score 1: 85.2±2.0%; score 2: 82.9±2.8%; MWU test; 0 vs. 1, p = 0.002; 1 vs. 2, p = 0.012; n = 360 scores; Fig 2G). This relation did not hold for food-restricted animals, probably because of the overall very low occurrence of >0 scores in this group (score 0: 85.5±1.6%; score 1: 84.9±1.7%; MWU test; 0 vs. 1, p = 0.30; n = 360 scores). In general, it is important to note that the differences between scores can be quite subtle, as is illustrated in Fig 2H and 2I, depicting a mouse with a score of 0 next to another one that had a total score of 2 (look/posture = 1; signs of dehydration = 1).

Continuous monitoring of physical activity in the home cage

While the discomfort score featured an instantaneous assessment of physical activity of the mice (activity and behavior), this could not be assessed without disturbing the mice in the first place. In order to measure activity of mice during the entire 24-hour cycle, we recorded the activity of each mouse in its home cage. Individual measurements were normalized to baseline activity as observed before the start of food/water restriction (see Methods; Fig 3A). These continuous readings were sensitive enough to measure the gradual adaptation to the reversed day/night cycle during the first seven days of the experiment (Fig 3B) and alterations to the day/night rhythm during the first two days after head-bar implantation surgery (Fig 3C).

Continuous home-cage activity recordings allowed us to monitor both the acute and long-term effects of restricted access to food or water. During the first few hours after restriction commenced, both experimental groups showed increased activity as compared to the non-implanted (non-restricted) group, which might be explained by the change into a novel cage (Fig 3D). On the following days, water-restricted animals showed a gradual decline in their daily activity, while food-restricted mice initially increased their home cage activity (Fig 3D). This initial increase in activity could indicate food-seeking/digging behaviors, before the animal learns that such efforts go unrewarded.

Across the entire duration of restriction, both food- and water-restricted mice showed reduced activity in the (active) daily light-off period (10:00–22:00, excluding the period during which training was typically done), as compared to their respective baseline levels before restriction had started (food-restricted: baseline = 0.62±0.16; training = 0.42±0.12; MWU test, p = 0.015; n = 6 mice; water-restricted: baseline = 0.47±0.04; training = 0.25±0.04; MWU test, p = 0.003; n = 6 mice; Fig 3E and 3F). This reduction in activity, relative to baseline activity, was not significantly different between food- and water-restricted mice (food-restricted: activity percentage of baseline = 68.0%±14.6%; water-restricted: activity percentage of baseline = 53.6%±5.9%; MWU test, p = 0.0641; n = 12 mice). Finally, in the post-restriction period, during which food and water was available *ad libitum* again, the average daily activity returned to levels that were comparable to the pre-training baseline (food-restricted: post-restriction = 0.73±0.19; water-restricted: post-restriction = 0.45±0.07; Fig 3F). Thus, by using continuous home-cage recordings we observed that food and water restriction induced a reversible reduction of overall activity levels that went undetected using the instantaneous scoring method.

Operant behavior and task-motivation

To compare how well the method of food and water restriction motivated mice to work for reward in a behavioral paradigm, we compared the total number of completed trials that mice

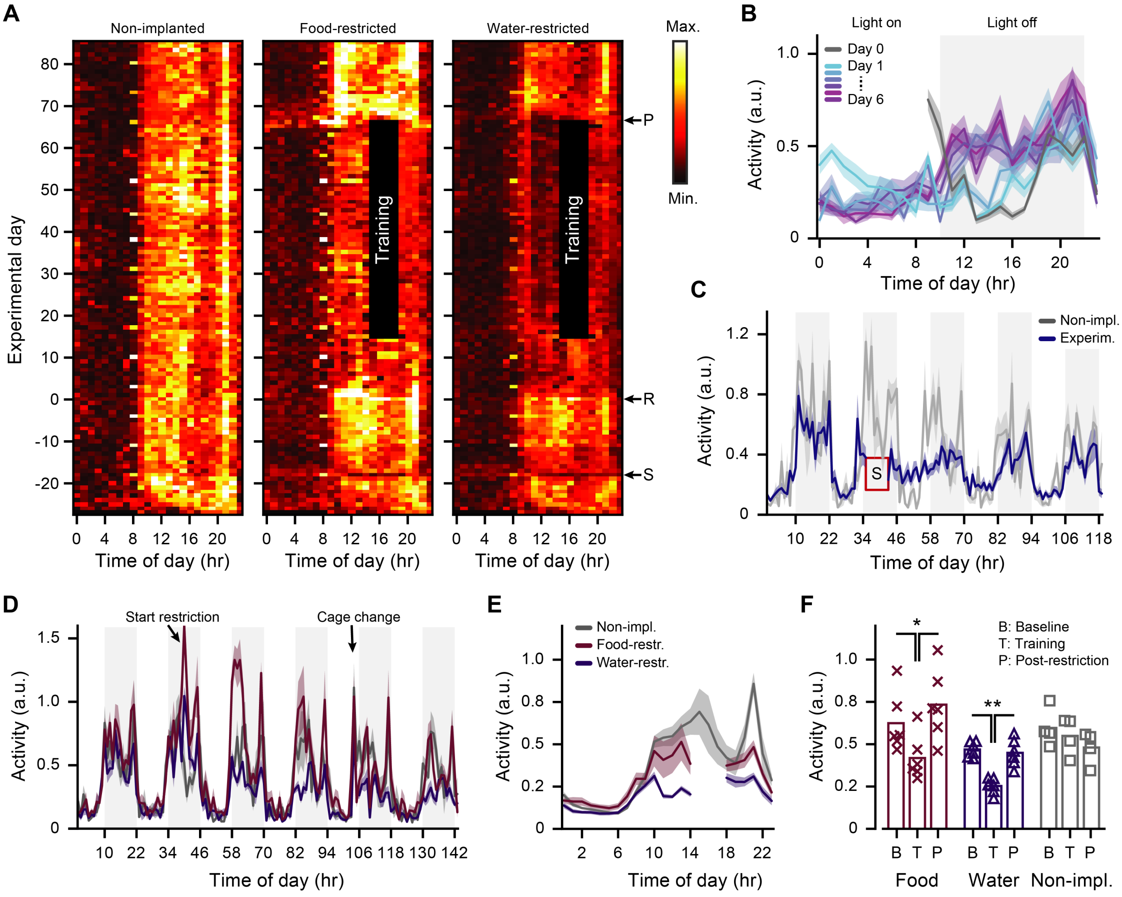


Fig 3. Continuous monitoring of physical activity in the home cage. A. Heat maps depicting baseline-normalized physical activity per hour (x-axis) throughout the days of the experiment (y-axis) as measured in the home cage, averaged across the non-implanted and experimental groups separately. Arrows: 'S' indicates the two days on which surgeries were performed; 'R' the day on which food or water restriction started; 'P' start of the post-training period (note that the food-restricted group received *ad libitum* access to food from two days before this post-training period). Cage changes can be identified as single bright data points, weekly reoccurring at 08:00. B. Hourly averaged (±SEM) home cage activity for the first seven days of adaptation to the reversed day/night cycle. C. Hourly averaged (±SEM) activity centered on the day of head bar implantation (blue, experimental group) or a matched day for animals that did not receive a head bar (gray, non-implanted group). D. Six days of average hourly home cage activity (mean±SEM), starting one day before onset of food or water restriction. E. Average (±SEM) 24hr home cage activity pattern throughout the entire period of training. Data of the experimental groups during the period of training (14:00–18:00) were left out. F. Mean (±SEM) home cage activity in the (active) light-off period (training period excluded). 'B': Baseline period, day -14 to 0. 'T': Training period, day 1 to 66. 'P': Post-restriction period, day 67 to 85. Crosses, triangle and squares indicate data points from individual mice (* MWU test, p<0.02; ** MWU test, p = 0.003).

https://doi.org/10.1371/journal.pone.0204066.g003

did in the pre-training stages (where every finished trial resulted in delivery of 8 μl soymilk or water). In pre-training stage 1, and (to a lesser extent) in pre-training stage 2, food-restricted mice executed significantly more trials compared to water-restricted animals (pre-training 1, # trials, food-restricted: 226±57; water-restricted: 45±21; MWU test, p = 0.003; pre-training 2, # trials, food-restricted: 237±84; water-restricted: 119±38; MWU test, p = 0.023; n = 12 mice; Fig 4F). As a direct consequence of this difference in total trial number, water-restricted mice required more pre-training stage 1 and pre-training stage 2 sessions to reach criterion compared to food-restricted animals. While we have no clear explanation for this, we noted that in subsequent experiments in our laboratory using water-restriction mice needed fewer pre-

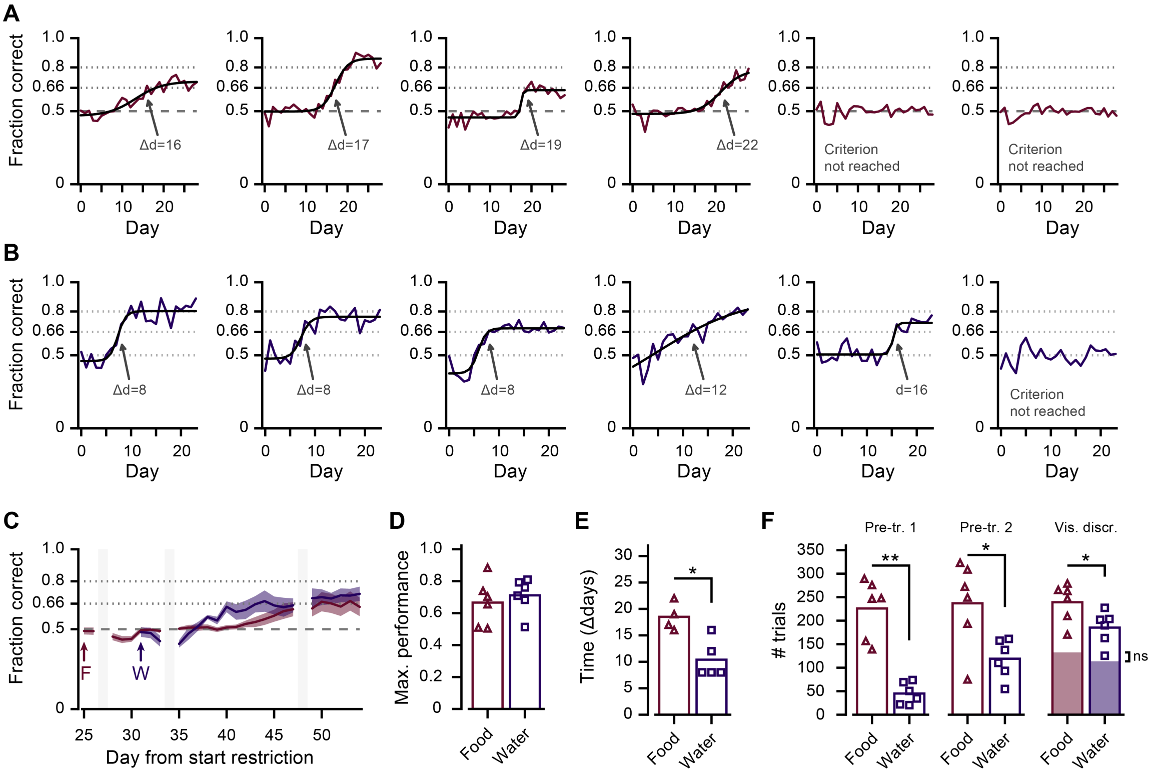


Fig 4. Visual discrimination in a head-fixed two-choice task. A, B. Learning curves of food (A) and water (B) restricted mice. Red and blue lines show the day-to-day performance for each animal, starting at the first day of visual discrimination learning. Black curve is a sigmoidal fit to data from animals that reached criterion (>66% correct). Gray arrows indicate the day on which mice reached criterion. C. Mean (±SEM) learning curve of all food- (red) and water- (blue) restricted mice in the overall experimental timeline. 'F' and 'W' indicate start days of training food- and water-restricted mice. Gray bars denote days without training. D. Maximum learning curve performance, determined by the sigmoidal fit in the time period during which mice were trained (as shown in A and B). E. Average number of days until criterion (>66% correct) was reached (MWU test, $^{*}p = 0.038$). F. Average number of trials that mice performed per day in each of the training stages ('Pre-tr 1': pre-training stage 1, 1 lick spout; 'Pre-tr 2': pre-training stage 2, 2 lick spouts; 'Vis. Discr.': visual discrimination task (training stage 3); MWU test, ns: not significant, $^{*}p = 0.023$, $^{**}p = 0.003$). The red/blue shaded area in bars of the visual discrimination stage indicates the fraction of rewarded (correct) trials.

https://doi.org/10.1371/journal.pone.0204066.g004

training sessions (three to five pre-training stage 1 and two to four pre-training stage 2 sessions). This indicates that potentially subtle changes in procedures, e.g. the experimenters becoming more experienced with the sub-millimeter positioning of the lick spouts, or only a single experimenter carrying out mouse handling and training, can reduce the number of pre-training sessions that water-restricted mice need.

In the final training stage (visual discrimination) the difference between the two groups was less pronounced, even-though food-restricted mice still performed a significantly larger number of trials compared to water-restricted mice (total # trials, food-restricted: 239±38; water-restricted: 185±33; MWU test, p = 0.023; Fig 4F). However, when only considering the trials in which mice made a correct choice, and thus received the soymilk or water reward, food- and water-restricted animals performed approximately equal numbers (Rewarded/correct # trials, Food-restricted: 132±26; Water-restricted: 114±28; MWU test, p = 0.149; n = 12 mice; Fig 4F).

The fact that water-restricted mice performed a lower number of trials throughout all training stages could indicate an overall lower motivation to work for water reward. Alternatively, water-restricted mice might satiate faster from water compared to how fast food-restricted mice satiate from soymilk, and therefore completed fewer trials. Anticipatory licking is a reward-oriented behavior that is related to task-motivation and that can be computed for individual trials [30]. By counting anticipatory licks in a 1 s period from stimulus onset until the response window started (Fig 1E) and averaging over only trials in which animals produced an operant response, we approximated task-motivation independent of satiation. This measure showed that during the final training sessions of the experiment, both food- and water-restricted mice made roughly 3 anticipatory licks per single trial (anticipatory # licks, last 10 sessions, food-restricted: 2.79±1.21; water-restricted: 2.93±0.42; MWU test, p = 0.189; n = 12 mice). However, during the first training sessions of the visual discrimination stage (stage 3), food-restricted mice systematically made fewer anticipatory licks compared to water-restricted mice (anticipatory # licks, first 10 sessions, food-restricted: 1.73±0.90; water-restricted: 3.30 ±0.91; MWU test, p = 0.015; n = 12 mice). This argues that the lower number of trials that water-restricted mice performed in each training stage did not reflect reduced motivation to lick for reward. Quite the opposite: the results rather suggest that water-restricted mice were even more motivated, but probably satiated faster compared to food-restricted animals.

To confirm that satiation is an important factor in task-motivation, we tested whether there was a correlation between the relative weight of a mouse and its behavioral drive. We found that the daily measurement of relative body weight significantly predicted the number of trials that a mouse would perform in the training session of the same day (Correlation of percentage body weight and total number of trials, z-scored per mouse; food-restricted: r = -0.22, p = 0.005; water-restricted: r = -0.47, p = $3.2 \cdot 10^{-10}$). In summary, food and water restriction can both be used to motivate animals in an operant task, but the total number of trials that mice perform depends on the restriction paradigm.

Operant learning and performance

To test for differences in operant learning, mice were trained to discriminate visual patterns in the two-choice head-fixed lick left/lick right task. Out of twelve mice, four food-restricted and five water-restricted mice reached the performance criterion of 66% correct trials on a given day (Fig 4A and 4B). For all mice, maximum learning curve performance was estimated from the fitted learning curve on the last day of training and did not differ between groups (Maximum of fitted learning curve, food-restricted: 0.67±0.13; water-restricted: 0.71±0.10; MWU test, p = 0.189; n = 12 mice; Fig 4C and 4D). However, water-restricted mice reached the criterion of 66% correct trials significantly faster compared to food-restricted mice (food-restricted, Δdays = 18.5±2.3; water-restricted, Δdays = 10.4±3.2; MWU test, p = 0.038; n = 9 mice; Fig 4C and 4E). This difference did not depend on the exact value of the threshold. A similar difference was observed with a higher threshold (70%, as in Guo et al., 2014; food-restricted, Δdays = 21.0±2.9; water-restricted, Δdays = 11.0±3.0; MWU test, p = 0.033; n = 9 mice) as well as with a lower threshold (60%; food-restricted, Δdays = 17.3±1.9; water-restricted, Δdays = 5.8 ±2.4; MWU test, p = 0.003; n = 10 mice). Also, using an altogether different method of quantifying whether the learning criterion was reached, the number of training sessions to reach the point of maximum steepness of the sigmoid fitted learning curve, we observed that water-restricted mice learned faster (food-restricted, Δdays = 18.0±2.8; water-restricted, Δdays = 7.8 ±4.8; MWU test, p = 0.046; n = 9 mice).

To investigate whether motivational state or satiation could explain the difference in speed of learning, we tested whether either the average relative weight-loss of a mouse, or the average number of anticipatory licks in 10 pre-learning sessions, predicted the number of sessions needed to reach learning criterion. However, neither variable correlated significantly with learning speed (correlation of percentage body weight and time to reach criterion, z-scored per condition: r = 0.11, p = 0.77; correlation of # of anticipatory licks and time to reach criterion, z-scored per condition: r = -0.54, p = 0.14; n = 9 mice). Additionally, we tested whether day-to-day fluctuations in relative body weight predicted task performance on the corresponding day (in mice that performed above criterion), which also did not correlate significantly in either the food- or water-restricted group (correlation of session-wise percentage body weight and performance, z-scored per mouse; food-restricted: r = -0.19, p = 0.28, n = 34 sessions; water-restricted: r = 0.11, p = 0.43, n = 58 sessions).

Another factor that may influence learning is general locomotor activity such as wheel-running in the home cage [31]. However, while food and water restriction both led to an overall reduction of home cage activity (Fig 3F), we observed that the most active mice actually took the longest to reach criterion (correlation of mean DVC activity and time to reach criterion, z-scored per condition: r = 0.82, p = 0.0069; n = 9 mice). In contrast to home-cage activity, we noted a large difference in the amount of running that the two groups of mice did during the visual discrimination task. Here, the water-restricted group ran about double the distance of the food-restricted group (distance ran per training session, food-restricted: 33±11 m; water-restricted: 66±24 m; MWU test, p = 0.023; n = 12 mice). Still, day-to-day differences in the amount of in-task running did not predict the performance on the visual discrimination task in either group (correlation of distance ran and performance, z-scored per mouse; food-restricted: r = 0.19, p = 0.28, n = 34 training sessions; water-restricted: r = 0.05, p = 0.72, n = 54 training sessions), and neither did the overall amount of in-task running predict the speed of learning (correlation in-task distance ran and time to reach criterion, z-scored per condition: r = -0.16, p = 0.67; n = 9 mice). Thus, parameters associated with motivational-state and physical activity provided a poor prediction of learning speed or task performance and do not likely explain the difference in time to reach learning criterion of food- and water-restricted mice.

Discussion

This study provides a detailed behavioral protocol for training mice in a fast and reliable way on a head-fixed two-alternative choice visual discrimination task. Our results show that most of the animals that were trained on the protocol learned discriminating visual stimuli within two or three weeks from the start of visual conditioning. An important aim of this study was to utilize the welfare- and behavioral read-outs of our training protocol to contrast two commonly used methods for motivating animals in head-fixed behavioral paradigms, i.e. food restriction with soymilk reward and water restriction with water reward. Using either method, the animals could be motivated to perform the task at or above criterion, without exceeding the 'mild' discomfort category, even for prolonged periods. However, we did observe specific differences in welfare assessment and in task performance, such as time to reach criterion and number of performed trials, which should be considered when selecting the restriction method.

Operant behavior

Throughout the training stages, there was a systematic difference in the number of trials that water- and food-restricted mice performed per session. In pre-training stage 1 and 2, food-restricted mice consumed larger volumes of soymilk than water-restricted mice consumed water. Furthermore, food-restricted animals proceeded faster through the pre-training stages than water-restricted mice. These differences might be explained by water-restricted mice satiating faster than food-restricted mice, since the 8-microliter water reward equaled on average 0.43% of the daily water intake in our experiment (1.93 ml), and the 8-microliter soymilk reward provided only 0.05% of the daily caloric food intake in this study (11.15 kcal). However, in pre-training 1 sessions water-restricted mice performed on average only 50 trials, which is approximately only 25% of their daily water ration. Possibly, water-restricted mice already satiate after 40–50 drops and only through experiencing multiple pre-training sessions learn to obtain more water than they acutely need. On the other hand, Guo et al. [9] observed that water-restricted mice performed more trials when sucrose was added to the water reward. Similarly, in our experiment we noted that water-restricted mice after reaching criterion performed more trials when provided with soymilk reward compared to the usual water reward (data not shown). Therefore, soymilk reward may have had additional motivating or appetitive aspects compared to plain water reward. Possibly, this is related to the nutrients and flavor that soymilk contains. Alternatively, it is conceivable that the smell of reward (soymilk) coming directly from the lick spouts made it easier for food-restricted (soymilk rewarded) mice to learn the initial behavior of licking for reward.

Despite water-restricted mice performing fewer trials per session, they were on average faster in reaching the learning criterion (independent of which exact criterion we used). Throughout the experiment, we aimed for keeping the motivational state of individual animals comparable by maintaining the relative weight of each animal at 85% of the *ad libitum* measured reference value. In addition, we excluded that the difference in learning speed correlated with parameters reflecting task-motivation in this study. One remaining explanation could be that, as described above, the 8-microliter water drop might have been subjectively perceived as a larger reward compared to a soymilk drop of the same volume, thus providing a greater learning incentive for water-restricted mice. Moreover, there are fundamental differences in the neural circuits that mediate hunger and thirst [22], asserting different effects on motivation and learning that could provide a stronger incentive for learning in one group compared to the other. Importantly, the speed of learning, maximum learning curve performance and success rate achieved using either restriction method in this study was similar to previously reported head-fixed operant conditioning paradigms, e.g. [6,9].

A final in-task difference between food- and water-restricted animals was the distance they ran on the Styrofoam ball during the period of behavioral training, with food-restricted mice running significantly less than water-restricted mice. While speculative, one possible explanation is that water-restricted mice are in a higher anticipation state during the task, as they receive relatively more of their daily water amount within-task compared to the relative caloric amount that food-restricted mice receive during the task, leading to hyperactive behavior [32]. Another explanation is that water-restricted mice spend more time running because they performed fewer trials in each training session and therefore had more time in which they were not task-engaged.

Welfare assessment

We aimed to facilitate the objective categorical distinction between methods of food and water restriction by maintaining mice in both conditions close to a body weight of 85% of their baseline reference. Still, there remained a limited amount of day-to-day and mouse-to-mouse variation of relative body weight within the protocol. These fluctuations correlated with the number of trials that mice performed in the behavioral task, showing that relative weight loss is a key factor in task-motivation. In case of water-restricted mice, relative weight loss also corresponded to the daily discomfort score (Fig 2G). One reason the absence of a correlation between relative weight loss and discomfort scores in food-restricted mice could be that this group had mostly discomfort scores of zero. Alternatively, it is conceivable that scores directly attributable to dehydration, such as skin tenting and fur appearance, were more sensitive or better observable compared to scores related to reduced food intake, such as the body condition score.

Although we only once observed reduced activity using the instantaneous scoring method, both food- and water-restricted mice showed less home-cage activity in the continuous home-cage recordings compared to their own baseline measurement. In food-restricted mice, the reduction in overall home-cage activity could reflect reduced food-seeking behavior. In water-restricted mice, the reduced home-cage activity might (in part) reflect a decrease in grooming behavior, which could be a factor contributing to their higher score on the parameter 'look/posture' in the overall health assessment. Indeed, water loss in the form of saliva used for grooming can account for up to one third of water loss in rodents that are not water-restricted [33]. It is possible that water-restricted mice conserve water by reducing the amount of grooming, leading to overall poorer fur appearance.

It should be noted that we did not assay the effect of food or water restriction on the mouse's physiology and neuronal circuitry, neither did we measure the effect of water restriction on food-intake behavior. In addition, the five scoring parameters may have differed in their sensitivity for detecting food- or water restriction associated discomfort. Therefore, we do not aim to draw conclusions from the differences in scores between restriction regimens observed in this study, but rather advise considering these results in the context of literature on food and water restriction procedures, e.g. [20,23,34,35]. Furthermore, the choice for setting a threshold at 85% of pre-restriction body weight is rather arbitrary. Other studies use different thresholds, either above or below 85%, and occasionally take into account the gradual increase in weight that would be observed in non-restricted mice, e.g. [6,36–38]. Still, these methods do not consider that there may be individual variation in how mice adapt to chronic water restriction [35,39]. Therefore, in our opinion, the best method would be to set the threshold for continuation of an experiment entailing food or water restriction using the measure of discomfort directly, as for instance described in Guo et al. [9], and monitor the relative weight of the animals as an indication, but not as threshold.

Practical considerations

In the last two decades, the mouse has gained increasing attention in neuroscience as a versatile research model that can be adopted for studying sensory processing, learning and memory, decision making and motor behavior under both healthy and diseased conditions. Our behavioral protocol and conditioning task for training head-fixed mice can be readily combined with *in vivo* recording techniques such as intracellular patch clamp recordings [40], two-photon microscopy [13], but also with newly developed techniques for single cell control of neuronal activity patterns [41]. The two-choice lick left / lick right task can be easily adapted to include other sensory modalities, or expanded for the study of higher cognitive functions, making it a useful tool for studying mouse behavior in general. In addition, the in-task differences we observed between food- and water-restricted animals can be exploited in order to suit the specific behavioral requirements. Finally, we showed that the use of a continuous home-cage monitoring system allows expanding the quantification of animal wellbeing to include an objective measure of overall activity, which allows observing light-cycle adaptation, post-surgery recovery and effects of food and water restriction without disturbing the animals. Behavioral paradigms will likely always require precise fine-tuning of a large, mostly un-documented parameter space. The methods and procedures described in this study are intended to guide this process to smoother convergence while improving animal wellbeing.

Supporting information

S1 File. Head bar holder design. These files contain the designs of the head bar and of the components necessary for building the head bar holder. The files were produced in

SolidWorks (Dassault Systèmes) and can also be viewed using the free program eDrawings (http://www.edrawingsviewer.com).
(ZIP)

S2 File. Lick spout holder design. This file contains the design of the 3D printable lick spout holder. It was produced and can be opened using the online service TinkerCat (https://www.tinkercad.com). The file can also be opened with the free program eDrawings (http://www.edrawingsviewer.com) as well as with most software delivered with 3D printers.
(STL)

Author Contributions

Conceptualization: Pieter M. Goltstein, Sandra Reinert, Annet Glas, Tobias Bonhoeffer, Mark Hübener.

Formal analysis: Pieter M. Goltstein.

Funding acquisition: Tobias Bonhoeffer, Mark Hübener.

Investigation: Pieter M. Goltstein, Sandra Reinert, Annet Glas, Mark Hübener.

Methodology: Pieter M. Goltstein, Sandra Reinert, Annet Glas.

Writing – original draft: Pieter M. Goltstein, Sandra Reinert, Annet Glas, Tobias Bonhoeffer, Mark Hübener.

Writing – review & editing: Pieter M. Goltstein, Sandra Reinert, Annet Glas, Tobias Bonhoeffer, Mark Hübener.

References

1. Skinner BF. The behavior of organisms: An experimental analysis. New York: Appleton-Century; 1938.
2. Stone CP, Darrow CW, Landis C, Heath LL. Studies in the dynamics of behavior. Chicago: University of Chicago Press; 1932.
3. O'Keefe J, Dostrovsky J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 1971; 34: 171–175. PMID: 5124915
4. Harvey CD, Collman F, Dombeck DA, Tank DW. Intracellular dynamics of hippocampal place cells during virtual navigation. Nature. 2009; 461: 941–946. https://doi.org/10.1038/nature08499 PMID: 19829374
5. O'Connor DH, Peron SP, Huber D, Svoboda K. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. Neuron. 2010; 67: 1048–1061. https://doi.org/10.1016/j.neuron.2010.08.026 PMID: 20869600
6. Histed MH, Carvalho LA, Maunsell JH. Psychophysical measurement of contrast sensitivity in the behaving mouse. J Neurophysiol. 2012; 107: 758–765. https://doi.org/10.1152/jn.00609.2011 PMID: 22049334
7. Sanders JI, Kepecs A. Choice ball: a response interface for two-choice psychometric discrimination in head-fixed mice. J Neurophysiol. 2012; 108: 3416–3423. https://doi.org/10.1152/jn.00669.2012 PMID: 23019000
8. Abraham NM, Guerin D, Bhaukaurally K, Carleton A. Similar odor discrimination behavior in head-restrained and freely moving mice. PLoS One. 2012; 7: e51789. https://doi.org/10.1371/journal.pone.0051789 PMID: 23272168
9. Guo ZV, Hires SA, Li N, O'Connor DH, Komiyama T, Ophir E, et al. Procedures for behavioral experiments in head-fixed mice. PLoS One. 2014; 9: e88678. https://doi.org/10.1371/journal.pone.0088678 PMID: 24520413
10. Poort J, Khan AG, Pachitariu M, Nemri A, Orsolic I, Krupic J, et al. Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. Neuron. 2015; 86: 1478–1490. https://doi.org/10.1016/j.neuron.2015.05.037 PMID: 26051421

11. Burgess CP, Lak A, Steinmetz NA, Zatka-Haas P, Bai Reddy C, Jacobs EAK, et al. High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice. Cell Rep. 2017; 20: 2513–2524. https://doi.org/10.1016/j.celrep.2017.08.047 PMID: 28877482

12. Aronov D, Tank DW. Engagement of Neural Circuits Underlying 2D Spatial Navigation in a Rodent Virtual Reality System. Neuron. 2014; 84: 442–456. https://doi.org/10.1016/j.neuron.2014.08.042 PMID: 25374363

13. Dombeck DA, Khabbaz AN, Collman F, Adelman TL, Tank DW. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. Neuron. 2007; 56: 43–57. https://doi.org/10.1016/j.neuron.2007.08.003 PMID: 17920014

14. Harvey CD, Coen P, Tank DW. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature. 2012; 484: 62–68. https://doi.org/10.1038/nature10918 PMID: 22419153

15. Keller GB, Bonhoeffer T, Hübener M. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. Neuron. 2012; 74: 809–815. https://doi.org/10.1016/j.neuron.2012.03.040 PMID: 22681686

16. Andermann ML, Kerlin AM, Reid RC. Chronic cellular imaging of mouse visual cortex during operant behavior and passive viewing. Front Cell Neurosci. 2010; 4: 3. https://doi.org/10.3389/fncel.2010.00003 PMID: 20407583

17. Makino H, Komiyama T. Learning enhances the relative impact of top-down processing in the visual cortex. Nat Neurosci. 2015; 18: 1116–1122. https://doi.org/10.1038/nn.4061 PMID: 26167904

18. Allen WE, DeNardo LA, Chen MZ, Liu CD, Loh KM, Fenno LE, et al. Thirst-associated preoptic neurons encode an aversive motivational drive. Science. 2017; 357: 1149–1155. https://doi.org/10.1126/science.aan6747 PMID: 28912243

19. Heiderstadt KM, McLaughlin RM, Wright DC, Walker SE, Gomez-Sanchez CE. The effect of chronic food and water restriction on open-field behaviour and serum corticosterone levels in rats. Lab Anim. 2000; 34: 20–28. https://doi.org/10.1258/002367700780578028 PMID: 10759363

20. Toth LA, Gardiner TW. Food and water restriction protocols: physiological and behavioral considerations. Contemp Top Lab Anim Sci. 2000; 39: 9–17.

21. Krashes MJ, Shah BP, Madara JC, Olson DP, Strochlic DE, Garfield AS, et al. An excitatory paraventricular nucleus to AgRP neuron circuit that drives hunger. Nature. 2014; 507: 238–242. https://doi.org/10.1038/nature12956 PMID: 24487620

22. Jourjine N, Mullaney BC, Mann K, Scott K. Coupled Sensing of Hunger and Thirst Signals Balances Sugar and Water Consumption. Cell. 2016; 166: 855–866. https://doi.org/10.1016/j.cell.2016.06.046 PMID: 27477513

23. Tucci V, Hardy A, Nolan PM. A comparison of physiological and behavioural parameters in C57BL/6J mice undergoing food or water restriction regimes. Behav Brain Res. 2006; 173: 22–29. https://doi.org/10.1016/j.bbr.2006.05.031 PMID: 16870275

24. Giles JM, Whitaker JW, Moy SS, Fletcher CA. Effect of Environmental Enrichment on Aggression in BALB/cJ and BALB/cByJ Mice Monitored by Using an Automated System. J Am Assoc Lab Anim Sci. 2018; https://doi.org/10.30802/AALAS-JAALAS-17-000122 [Epub ahead of print]. PMID: 29669621

25. Ullman-Culleré MH, Foltz CJ. Body condition scoring: a rapid and accurate method for assessing health status in mice. Lab Anim Sci. 1999; 49: 319–323. PMID: 10403450

26. Hölscher C, Schnee A, Dahmen H, Setia L, Mallot HA. Rats are able to navigate in virtual environments. J Exp Biol. 2005; 208: 561–569. https://doi.org/10.1242/jeb.01371 PMID: 15671344

27. Scott BB, Brody CD, Tank DW. Cellular resolution functional imaging in behaving rats using voluntary head restraint. Neuron. 2013; 80: 371–384. https://doi.org/10.1016/j.neuron.2013.08.002 PMID: 24055015

28. Weijnen JA. Lick sensors as tools in behavioral and neuroscience research. Physiol Behav. 1989; 46: 923–928. PMID: 2634256

29. Slotnick B. A simple 2-transistor touch or lick detector circuit. J Exp Anal Behav. 2009; 91: 253–255. https://doi.org/10.1901/jeab.2009.91-253 PMID: 19794837

30. DeBold RC, Miller NE, Jensen DD. Effect of Strength of Drive Determined by a New Technique for Appetitive Classical Conditioning of Rats. J Comp Physiol Psychol. 1965; 59: 102–108. PMID: 14282384

31. Li H, Liang A, Guan F, Fan R, Chi L, Yang B. Regular treadmill running improves spatial learning and memory performance in young mice through increased hippocampal neurogenesis and decreased stress. Brain Res. 2013; 1531: 1–8. https://doi.org/10.1016/j.brainres.2013.07.041 PMID: 23916669

32. Sherwin CM. Voluntary wheel running: a review and novel interpretation. Anim Behav. 1998; 56: 11–27. https://doi.org/10.1006/anbe.1998.0836 PMID: 9710457

33. Ritter RC, Epstein AN. Saliva lost by grooming: a major item in the rat's water economy. Behav Biol. 1974; 11: 581–585. PMID: 4415674

34. Haines H, Ciskowski C, Harms V. Acclimation to chronic water restriction in the wild house mouse Mus musculus. Physiological Zoology. 1973; 46: 110–128.

35. Rowland NE. Food or fluid restriction in common laboratory animals: balancing welfare considerations with scientific inquiry. Comp Med. 2007; 57: 149–160. PMID: 17536615

36. Busse L, Ayaz A, Dhruv NT, Katzner S, Saleem AB, Schölvinck ML, et al. The detection of visual contrast in the behaving mouse. J Neurosci. 2011; 31: 11351–11361. https://doi.org/10.1523/JNEUROSCI.6689-10.2011 PMID: 21813694

37. Montijn JS, Goltstein PM, Pennartz CM. Mouse V1 population correlates of visual detection rely on heterogeneity within neuronal response patterns. Elife. 2015; 4: e10163. https://doi.org/10.7554/eLife.10163 PMID: 26646184

38. Jurjut O, Georgieva P, Busse L, Katzner S. Learning enhances sensory processing in mouse V1 before improving behavior. J Neurosci. 2017; 37: 6460–6474. https://doi.org/10.1523/JNEUROSCI.3485-16.2017 PMID: 28559381

39. Bekkevold CM, Robertson KL, Reinhard MK, Battles AH, Rowland NE. Dehydration parameters and standards for laboratory mice. J Am Assoc Lab Anim Sci. 2013; 52: 233–239. PMID: 23849404

40. Komai S, Denk W, Osten P, Brecht M, Margrie TW. Two-photon targeted patching (TPTP) in vivo. Nat Protoc. 2006; 1: 647–652. https://doi.org/10.1038/nprot.2006.100 PMID: 17406293

41. Mardinly AR, Oldenburg IA, Pégard NC, Sridharan S, Lyall EH, Chesnov K, et al. Precise multimodal optical control of neural ensemble activity. Nat Neurosci. 2018; 21: 881–893. https://doi.org/10.1038/s41593-018-0139-8 PMID: 29713079

3 | Mouse prefrontal cortex represents learned rules for categorization

Mouse prefrontal cortex represents learned rules for categorization

https://doi.org/10.1038/s41586-021-03452-z

Received: 25 November 2019

Accepted: 12 March 2021

Published online: 21 April 2021

Open access

Check for updates

Sandra Reinert[1,2], Mark Hübener[1], Tobias Bonhoeffer[1] & Pieter M. Goltstein[1]✉

The ability to categorize sensory stimuli is crucial for an animal's survival in a complex environment. Memorizing categories instead of individual exemplars enables greater behavioural flexibility and is computationally advantageous. Neurons that show category selectivity have been found in several areas of the mammalian neocortex[1–4], but the prefrontal cortex seems to have a prominent role[4,5] in this context. Specifically, in primates that are extensively trained on a categorization task, neurons in the prefrontal cortex rapidly and flexibly represent learned categories[6,7]. However, how these representations first emerge in naive animals remains unexplored, leaving it unclear whether flexible representations are gradually built up as part of semantic memory or assigned more or less instantly during task execution[8,9]. Here we investigate the formation of a neuronal category representation throughout the entire learning process by repeatedly imaging individual cells in the mouse medial prefrontal cortex. We show that mice readily learn rule-based categorization and generalize to novel stimuli. Over the course of learning, neurons in the prefrontal cortex display distinct dynamics in acquiring category selectivity and are differentially engaged during a later switch in rules. A subset of neurons selectively and uniquely respond to categories and reflect generalization behaviour. Thus, a category representation in the mouse prefrontal cortex is gradually acquired during learning rather than recruited ad hoc. This gradual process suggests that neurons in the medial prefrontal cortex are part of a specific semantic memory for visual categories.

We trained head-fixed mice ($n = 11$) in a 'Go'/'NoGo' rule-based categorization task (Fig. 1a, b) to sort visual stimuli into two groups. Stimuli were 36 sinusoidal gratings, each with a specific combination of two stimulus features: orientation and spatial frequency. At any time, one rule determined the relevant feature for categorization (that is, the active rule; for example, assigning category identity of a stimulus based on orientation)[10,11] (Extended Data Fig. 1). First, mice learned to discriminate two exemplar stimuli according to the active rule. All mice achieved more than 66% correct Go choices within 10 to 40 sessions, showing considerable variability in the rate of learning. We then introduced categories by stepwise addition of stimuli to the task, up to a set of 18 different gratings that varied along both feature dimensions, orientation and spatial frequency (Extended Data Fig. 1b, c). Mice integrated the newly introduced stimuli within 1 to 2 sessions and they maintained a sensitivity index d' of >1 (Fig. 1c, d, Extended Data Figs. 1d, 2).

Mice learn to categorize visual stimuli

To determine whether mice had indeed learned categorization, we tested a characteristic feature of category learning, rapid generalization to novel stimuli[10–13]. Mice were presented with 18 novel grating stimuli in addition to the 18 experienced ones. All mice were able to generalize the learned rule to novel stimuli upon their first presentation (time point 5, T5) (Fig. 1d, Extended Data Fig. 3a), performing equally well on novel and experienced stimuli (Fig. 1e, Extended Data Fig. 3b).

Because rule-switching is key to rule-based categorization[11,14,15], our stimulus set was designed to allow for testing this aspect. Thus, after learning the first rule, mice were required to group the same stimuli into new categories according to a new rule, by making the previously irrelevant stimulus feature (for example, spatial frequency) relevant and the relevant one (for example, orientation) irrelevant. All mice learned to discriminate two exemplar stimuli for rule 2 considerably faster than during initial learning (Fig. 1f, Extended Data Fig. 1e–h). After the mice had learned to categorize a set of 18 stimuli according to rule 2, they were able to generalize to the 18 stimuli they had so far experienced only with rule 1 (Fig. 1g, h, Extended Data Fig. 3c–f). We tested whether there were any residual effects of the former rule on the categorization behaviour of the mice by comparing the influence of each stimulus feature (Fig. 1i) on the choices of the mice before learning (time point T1) and after learning each rule (T5 and T8). Untrained mice showed no effect of either stimulus feature on categorization (Fig. 1j, left). Trained mice only based categorization on the stimulus feature relevant to the active rule; the irrelevant feature showed no effect (Fig. 1j, middle, right; for a detailed analysis of categorization near the boundary see Extended

[1] Max Planck Institute of Neurobiology, Martinsried, Germany. [2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Martinsried, Germany. ✉e-mail: goltstein@neuro.mpg.de

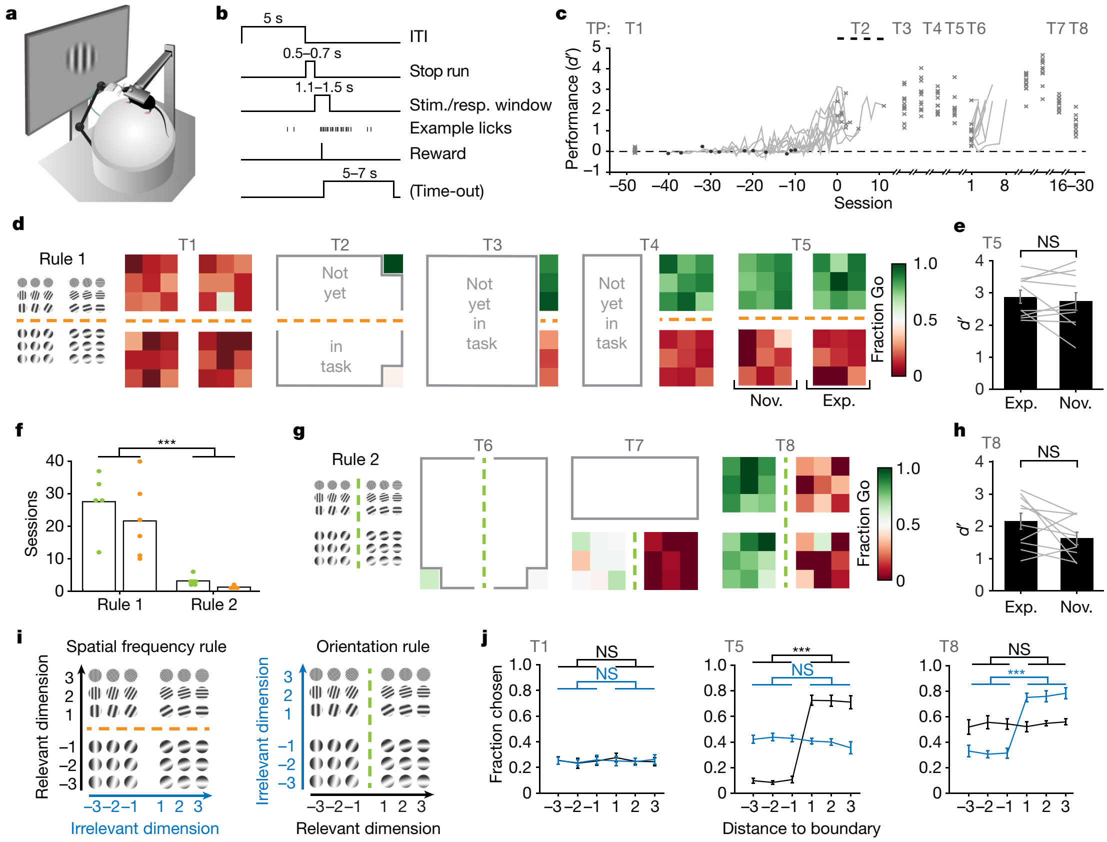


Fig. 1 | Mice learn rules to categorize visual stimuli and generalize to novel stimuli. a, Schematic of behavioural training setup. **b**, Schematic of trial structure in the Go/NoGo task. ITI, inter-trial interval; Stim./resp., stimulus presentation/response window. **c**, Performance (d') of 11 mice in each training session. Individual traces aligned to criterion (66% of correct trials). The dashed line indicates chance level ($d' = 0$). Crosses denote sessions with two-photon imaging (T1–T8). The spread in performance after T2 is due to day-to-day variability rather than mouse-to-mouse variability. TP, time point. **d**, Fraction of Go choices per stimulus of an example mouse at each time point (of two-photon imaging) until the presentation of all 36 stimuli of rule 1 (generalization; T5). **e**, Performance (d') for rule 1 (T5), for experienced (Exp.) compared to novel (Nov.) stimuli. $P = 0.50$, two-tailed paired-samples t-test ($n = 11$ mice). Grey lines denote individual mice. Data are mean ± s.e.m. **f**, Number of training sessions until criterion (66% correct, exemplar stimuli). Bars indicate mean across mice, dots are individual mice (green denotes the orientation rule; orange denotes spatial frequency rule). Rule 2 is learned significantly faster than rule 1. $P = 9.77 \times 10^{-4}$, two-tailed Wilcoxon matched-pairs signed-rank (WMPSR) test ($n = 11$ mice). **g**, As in **d**, for rule 2 of the same mouse. **h**, As in **e** for rule 2 (T8). d' did not differ significantly between novel stimuli and stimuli experienced with rule 2. $P = 0.09$, two-tailed paired-samples t-test ($n = 10$ mice). **i**, Schematics specifying the distance of stimuli to the boundary. **j**, Psychometric curves showing the fraction of Go choices along the relevant (black) and irrelevant (blue) dimension of rule 1 at T1, T5 and T8. Left: $P_{\mathrm{relevant\,T1}} = 0.36$, $P_{\mathrm{irrelevant\,T1}} = 0.77$; middle: $^{***}P_{\mathrm{relevant\,T5}} = 1.73 \times 10^{-6}$, $P_{\mathrm{irrelevant\,T5}} = 0.09$; right: $P_{\mathrm{relevant\,T5}} = 0.73$, $^{***}P_{\mathrm{irrelevant\,T5\,(relevant\,T8)}} = 1.73 \times 10^{-6}$, two-tailed WMPSR test, Bonferroni-corrected for two comparisons ($n = 10$ mice). Categorization performance was not affected by the order in which mice were trained on orientation and spatial frequency rules. Data are mean ± s.e.m. across mice; for individual mice, see Extended Data Fig. 2. NS, not significant.

Data Fig. 3g–l). In summary, all mice learned discriminating categories on the basis of two different rules, and they generalized these rules to novel stimuli, probably by selectively attending[16] to the relevant stimulus feature. Having established this training paradigm, we began tracking neuronal correlates of rule-based categories throughout learning.

mPFC contains category-selective neurons

The prefrontal cortex (PFC) in primates and rodents is important for cognitive functions such as categorization[6,7,16,17], rule learning[1,18,19] and cognitive flexibility[20–22], even though the functional and anatomical analogy of this region across species is still debated[23–28]. Earlier studies in primates have described individual PFC neurons coding for visual categories[3,6,7]. Encouraged by this finding, we tested whether the mouse medial PFC (mPFC) contained neurons that reflected the ability of the mouse to categorize visual stimuli as described above. To this end, we chronically monitored neuronal activity in cortical layer 2/3 using two-photon calcium imaging through a microprism implant inserted between the two hemispheres, which enabled optical access to mPFC[29] (Fig. 2a–c, Extended Data Fig. 4). We measured neuronal activity of individual cells while the mice performed the task (d' ranging from 0.7 to 3.6; for imaging time points and precise trial structure, see Fig. 1b, c, Extended Data Fig. 1a). In naive mice (time point T1), mPFC neurons did not respond to visual stimuli (Fig. 2b, d, Extended Data Fig. 5), but some of these initially non-selective cells clearly showed category selectivity after learning (T5, rule 1) (Fig. 2c, e, Extended Data Fig. 5; neural correlates of other task-related aspects are described below).

We quantified the category selectivity of individual cells using the category-tuning index (CTI)[30], with values close to 1 indicating strong differences in activity between, but not within, categories, and a value of 0 indicating no difference in the firing rate between and within categories. We defined neurons with a CTI value above 0.1 as category-selective

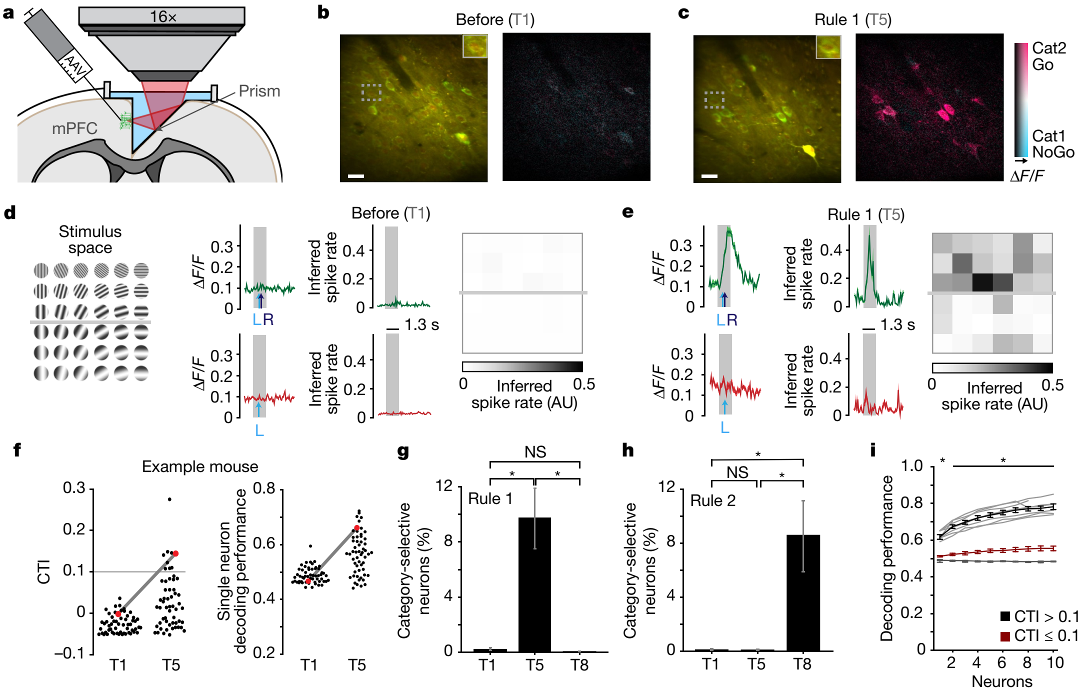


Fig. 2 | Single neurons in the mouse mPFC develop category-selective responses. a, Schematic depicting virus injection into the mPFC and two-photon imaging through a prism implant between the hemispheres, adapted from ref. [29]. AAV, adeno-associated virus. **b**, Example field of view before learning (T1). Left, pseudo-coloured GCaMP6m (green) and mRuby2 (red) fluorescence. Right, hue, saturation and lightness (HLS) map. Hue: preferred category; lightness: response amplitude; saturation: category selectivity. Scale bar, 30 µm. **c**, As in **b**, after learning (T5). **d**, Left, stimulus space. Middle, average response to stimuli of the cell highlighted in **b** (Δ*F*/*F*, green: Go category, red: NoGo category). L, average time of first lick. R, average time of reward delivery. Right, inferred spike rate. AU, arbitrary units. **e**, As in **d**, after learning, showing selective responsiveness to Go category stimuli. **f**, Left, CTI of all cells in the example field of view before and after learning. Red, example cell in **b**. Grey line denotes the threshold used for further analyses (CTI > 0.1). Right, cross-validated performance of a Bayesian decoder predicting the category of the presented stimulus. CTI correlates with decoding performance. $P = 2.2 \times 10^{-10}$, rho = 0.41, Spearman's correlation (n = 213 category-selective neurons, CTI > 0.1). **g**, Percentage of category-selective cells for rule 1. T1, before learning, rule 1 active. T5, after learning, rule 1. T8, after learning, rule 2. $P_{\mathrm{T1-T5}} = 0.006$, $P_{\mathrm{T1-T8}} = 0.25$, $P_{\mathrm{T5-T8}} = 0.004$, two-tailed WMPSR test, Bonferroni-corrected for three comparisons ($n_{\mathrm{mice}} = 10$, $n_{\mathrm{neurons}} = 2{,}306$). **h**, As in **g**, for rule 2. $P_{\mathrm{T1-T5}} = 0.75$, $P_{\mathrm{T1-T8}} = 0.004$, $P_{\mathrm{T5-T8}} = 0.004$, two-tailed WMPSR test, Bonferroni-corrected for three comparisons ($n = 10$ mice). **i**, Bayesian decoding performance as in **f**, for all mice. Data are shown separately for populations of low (red) and high (black) CTI cells. Light grey denotes individual mice; dark grey denotes average performance after shuffling stimulus categories. $P_{\mathrm{1\,neuron}} = 0.005$, $P_{\mathrm{2-8\,neurons}} = 0.01$, two-tailed WMPSR test, high versus low CTI cells, Bonferroni-corrected for two comparisons ($n_{\mathrm{1\,neuron}} = 10$ mice, $n_{\mathrm{2-8\,neurons}} = 8$ mice). Data are mean ± s.e.m. across mice; for individual mice see Extended Data Figs. 5, 6.

(Methods), and verified that these cells reliably encoded categories using cross-validated Bayesian decoding (Fig. 2f, i). In naive mice, hardly any cells exceeded this threshold, whereas after learning, a substantial fraction of neurons in the mPFC showed category selectivity (before: 0.03% ± 0.03%, after: 9.8% ± 2.2% (mean ± s.e.m.)) (Fig. 2g, Extended Data Fig. 6a).

After having learned the rule-switch, a similar fraction of cells showed selectivity for the new categories, whereas selectivity for the old, now irrelevant categories ceased (rule 1: 0.07% ± 0.05%, rule 2: 8.6% ± 2.8%) (Fig. 2h, Extended Data Fig. 6a).

To convert an internal category representation into a motor decision, it would be sufficient for cells in the mPFC to show selectivity for only one category[31]. However, we observed two types of neuron—one that represented rewarded stimuli (Go preferring: 73% of all category-selective cells at T5 and 65% at T8) and the other non-rewarded stimuli (NoGo preferring: 27% at T5 and 35% at T8). Thus, cells in the mouse mPFC develop flexible representations of rule-based categories over the course of learning.

Category selectivity emerges over time

Our chronic recording approach allowed us to ask whether the cells that coded for learned categories in rule 2 were the same ones that had represented categories in rule 1. Although many cells that were category-selective for rule 1 were less selective for rule 2, a subset of neurons remained category-selective throughout (Fig. 3a, Extended Data Fig. 7a, b). We found that, on average, the Go category-selective neurons remapped their responses to the new Go category—that is, after the rule-switch, they responded to a different set of visual stimuli. By contrast, the NoGo category-selective cells did not remap (Fig. 3a, Extended Data Fig. 7c, d). They lost their selectivity after the rule-switch, and a new set of cells became NoGo category-selective for the newly defined categories. Similarly, the Go category-selective cells observed after the rule-switch showed previous selectivity to the first rule, whereas rule 2 NoGo category-selective neurons did not show any selectivity before the rule-switch on average (Fig. 3b). In line with this, we observed that the Go category-selective populations for each rule overlapped more than expected by chance (Methods), in contrast to NoGo category-selective populations (Extended Data Fig. 6b–d). Notably, neurons were less likely than chance to switch their preference from Go to NoGo and vice versa (Extended Data Fig. 6b).

It is currently debated whether such flexible representations in the PFC are gradually built up during learning—that is, are part of the memory of the learned categories—or whether they are instantaneously

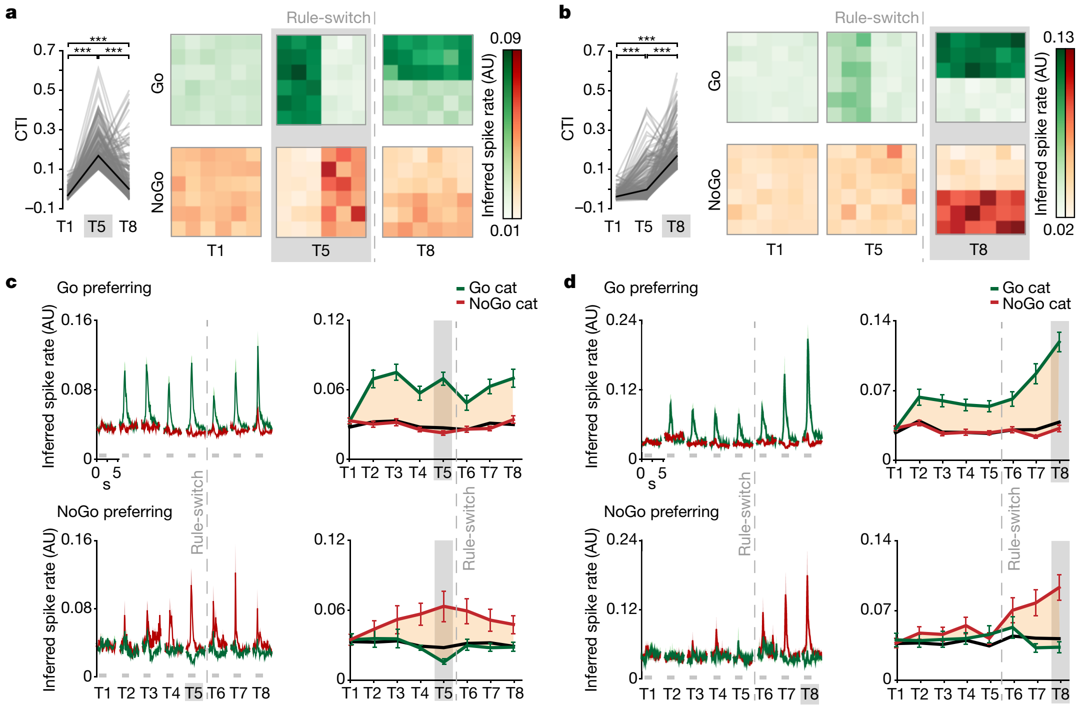


Fig. 3 | Two populations of category-selective neurons show different dynamics during a rule-switch. a, Left, CTI of all category-selective neurons identified at T5 (grey highlight), shown for time points T1, T5 and T8. $P_{\text{T1–T5}} = 1.1 \times 10^{-36}$, $P_{\text{T1–T8}} = 1.6 \times 10^{-14}$, $P_{\text{T5–T8}} = 6.9 \times 10^{-27}$, two-tailed WMPSR test, Bonferroni-corrected for three comparisons ($n = 213$ cells). Black line denotes the mean. Right, average inferred spike rate per stimulus of Go and NoGo category-selective cells at T1, T5 and T8 ($n_{\text{Go}} = 156$ cells; $n_{\text{NoGo}} = 57$ cells) **b**, As in **a**, but for category-selective cells defined at T8. $P_{\text{T1–T5}} = 4.2 \times 10^{-18}$, $P_{\text{T1–T8}} = 2.9 \times 10^{-33}$, $P_{\text{T5–T8}} = 1.1 \times 10^{-27}$, two-tailed WMPSR test, Bonferroni-corrected for three comparisons ($n = 192$ cells; $n_{\text{Go}} = 122$ cells; $n_{\text{NoGo}} = 70$ cells). **c**, Left, inferred spike rate of all Go (top) and NoGo (bottom) category-selective cells, identified at T5 (grey highlight), during trials of all Go (green) and NoGo (red) category stimuli at T1–T8. Grey denotes stimulus presentation. Data are mean ± s.e.m., across cells. Right, inferred spike rate during stimulus presentation of all Go (top) and NoGo (bottom) category-selective cells. Green denotes Go category, red denotes NoGo category, orange area denotes the difference. Black denotes the mean inferred spike rate in the pre-stimulus period. Data are mean ± s.e.m., across cells. **d**, As in **c**, for category-selective cells defined at T8.

assigned during the task to represent anything that becomes relevant to the animal[7–9]. This question can be answered only by monitoring neurons throughout the learning process, starting from a naive animal. We took advantage of the fact that our mice had never been trained on a categorization task and we followed the development of category-selective responses of individual neurons over the entire time course of rule-based category learning (Fig. 3c). Focusing on the period over which selectivity emerged, we observed a marked difference between the time courses that the Go and NoGo category-selective neurons followed. On average, the Go category-selective cells showed large, stable responses for the Go category, early on after presentation of the initial category stimuli in an ad hoc fashion (T2–T5) (Fig. 3c, Extended Data Fig. 7e). By contrast, the NoGo category-selective cells only gradually developed selectivity with increasing categorization demand of the task (T4, T5) (Fig. 3c, Extended Data Fig. 7f). After the rule-switch, the Go category-selective cells on average switched their stimulus selectivity, thereby retaining category selectivity. Former NoGo category-selective cells gradually lost selectivity, whereas a new, independent population of NoGo category-selective neurons gained selectivity (Fig. 3c, d). Notably, after the rule-switch, Go category-selective neurons showed increased Go responsiveness beyond a stable level of Go selectivity during earlier training (Fig. 3d).

A possible explanation for the different time courses could be that various task-relevant components differentially contribute to the average selectivity. It is well established that—beyond the category selectivity we observed—the mPFC contains representations of choice and reward[32–35]. In our paradigm, choice and reward associations are learned earlier than categories, and stay constant through the rule-switch. Therefore, neurons selective for choice and reward are expected to show a different time course than neurons selective for stimulus category (Extended Data Fig. 7g). We identified individual neurons that acquire selectivity early-on during task learning as well as neurons that develop selectivity more gradually, with increasing categorization demand (Extended Data Fig. 7h–k). In line with their average (Fig. 3c, d), most NoGo-preferring neurons followed the gradual time course, reflecting acquisition of the respective category rule, whereas Go-preferring neurons followed either of the time courses (Extended Data Fig. 7h–k). Thus, neurons that prefer the Go category were modulated by category, as well as by the earlier learned reward and choice associations (Extended Data Fig. 8).

To disentangle how stimulus category, choice and reward affected the trial-by-trial responses of category-selective neurons, we used linear regression to determine their individual contributions (Extended Data Fig. 9a). Although choice selectivity did not directly explain CTI (Extended Data Fig. 9b, c), the activity pattern of Go category-selective cells showed significant modulation by multiple factors, stimulus category, choice and reward (Extended Data Fig. 9d). By contrast, the responses of NoGo category-selective cells were only significantly influenced by category identity (Extended Data Fig. 9d). We performed hierarchical clustering to explore the entire task-responsive neuronal population in the mPFC including category-selective cells and found

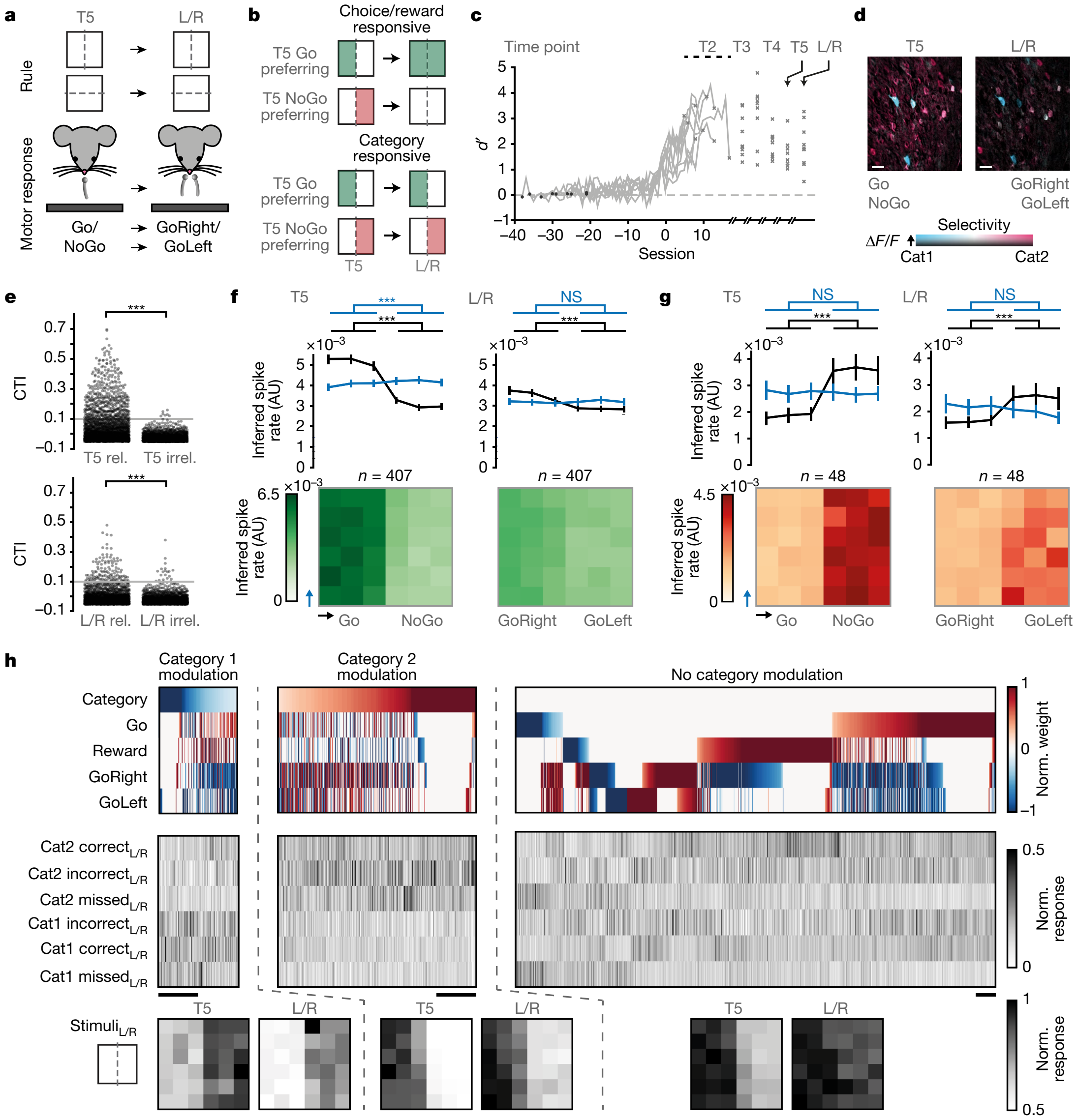


Fig. 4 | Mouse mPFC contains uniquely category-modulated neurons. **a**, Schematic depicting the change in task from Go/NoGo (T5) to left/right choice (L/R). The motor response changed from Go to GoRight, and from NoGo to GoLeft. The category rule remained the same. **b**, Possible changes in neuronal responses between T5 and left/right choice. Top, choice/reward-selective neurons. Bottom, uniquely category-selective neurons. **c**, *d′* throughout category learning and the change in task, aligned to criterion (>66% correct, *n* = 9 mice). **d**, Example HLS maps before (T5) and after (L/R) the task change. Scale bars, 30 µm. Hue: preferred category; lightness: response amplitude; saturation: category selectivity. **e**, CTI of all recorded neurons, calculated using either the relevant or the irrelevant rule, before (T5) and after (L/R) the task change. T5: $P = 1.3 \times 10^{-161}$, L/R: $P = 1.3 \times 10^{-24}$, two-tailed WMPSR test ($n = 2{,}389$). Grey lines denote CTI = 0.1. **f**, Top, inferred spike rate for stimuli ordered along the relevant dimension (black), or the irrelevant dimension (blue) across all Go category-selective neurons (defined at T5, CTI > 0.1). $P_{\mathrm{T5\,rel}} = 6.5 \times 10^{-28}$, $P_{\mathrm{T5\,irrel}} = 1.6 \times 10^{-5}$, $P_{\mathrm{L/R\,rel}} = 1.6 \times 10^{-19}$, $P_{\mathrm{L/R\,irrel}} = 0.38$, two-tailed WMPSR test ($n = 407$). Data are mean ± s.e.m. Bottom, mean activity per stimulus. **g**, As in **f**, for NoGo category-selective neurons. $P_{\mathrm{T5\,rel}} = 1.9 \times 10^{-4}$, $P_{\mathrm{T5\,irrel}} = 0.45$, $P_{\mathrm{L/R\,rel}} = 7.7 \times 10^{-4}$, $P_{\mathrm{L/R\,irrel}} = 0.21$, two-tailed WMPSR ($n = 48$). **h**, Predictor weights and response amplitudes of significantly modulated neurons. $P < 0.05$, at least one predictor, *F*-statistic (1,904 neurons). Scale bars, 50 neurons. Top row, normalized predictor weights for each neuron. Left, neurons with a negative category weight (category 1, NoGo/GoLeft). Middle, neurons with a positive category weight (category 2, Go/GoRight). Right, no significant category modulation. Middle row, average (normalized) activity in correct, incorrect and missed trials of categories 1 and 2 separately. Bottom row, per group, the mean normalized response to all presented stimuli at T5 (left) and L/R (right).

clusters of mPFC neurons that were predominantly modulated by a single parameter—that is, category, choice (lick) and reward (Extended Data Fig. 9e–i, cluster number 1, 2 and 3, respectively). In addition, there were also clusters of neurons modulated by specific combinations of task parameters (Extended Data Fig. 9i, clusters number 4, 5 and 9).

These results are in line with recent studies in primates and mice, reporting mixed selectivity of neurons in the PFC after animals learned cognitive tasks[21,36–38]. In summary, the mouse mPFC contains neurons modulated by a single parameter (such as category) and neurons that show mixed-selective responses.

Category tuning generalizes across tasks

Because the activity of many mPFC neurons, including category-selective neurons, correlated with combinations of stimulus category, choice and reward, we aimed to experimentally determine the unique category-selective component. Exclusively category-modulated neurons can be revealed by experimental decoupling of the presented category and the associated motor response[39]. Within the framework of our rule-based categorization paradigm, we achieved this by initially training mice to categorize in the Go/NoGo task (as before), and then changing the task to a left/right choice design (Fig. 4a). As a consequence, the previous Go (lick) category changed into a 'GoRight' (lick right) category, and the previous NoGo (no lick) category was now also rewarded if the mouse made a 'GoLeft' (lick left) response. In this experiment, neurons that were category-selective in the Go/NoGo task could either retain their category selectivity in the left/right choice task (indicating that they are genuinely category-selective), or change their response pattern, reflecting selectivity rather for motor planning, choice or associated reward (Fig. 4b).

We first trained nine mice to categorize visual stimuli according to either the spatial frequency or the orientation rule (the task was identical to that in Fig. 1 and Extended Data Fig. 1, up to the generalization test T5; Fig. 4c). After session T5, we changed the behavioural setup by replacing the single centred lick spout with two laterally placed lick spouts (left/right choice paradigm). The mice quickly adapted to the change and within the first four trials also responded with licks to the previous NoGo category (now GoLeft; note that the ratio between the left and right licks varied throughout the session). Although the mice did not specifically target their first licks to the correct spout, they performed a similar number of licks on both lick spouts and obtained a similar amount of rewards for both categories.

We found a significant proportion of category-selective neurons before (T5) and after the task change (left/right; threshold of CTI > 0.1 according to the relevant rule) (Fig. 4d, e). On average, category-selective cells identified at T5 discriminated the stimulus categories also after the task change (Fig. 4d, f, g, Extended Data Fig. 10a–g), although their selectivity decreased. The left/right choice task allowed us to compare trials with different stimulus categories in the absence of choice and reward (missed trials). Neurons that were initially selective for the Go category remained selective for the same stimulus category. Likewise, initial NoGo category-selective neurons, remained only responsive to stimuli of the previous NoGo category (Extended Data Fig. 10h, i).

However, the overall decrease in selectivity after the task change indicated that also choice-and reward-selective neurons were identified as 'category-selective' in the Go/NoGo task (Fig. 4b). Because the left/right choice task changed how reward and motor contingencies mapped onto the stimulus space, but did not change the mapping of category identity, we were able to use a regression model to disambiguate these contributions. Only neurons that remained category-selective across the task change will be significantly fitted by the Category predictor. Apparent category-selective neurons—that is, choice- and reward-modulated neurons, will be better predicted by the Go and Reward predictors. This analysis showed that mouse mPFC neurons represent categories in conjunction with reward and choice. Most importantly, it also revealed a set of uniquely category-modulated neurons in the mPFC (4.3%) (Fig. 4h, Extended Data Fig. 10j).

Recent work has shown the influence of uninstructed behaviours, such as whisking and eye movements, on neuronal response variability in operant tasks[40]. If such behaviours correlated with the category identity of the presented stimuli, they could lead to apparent category selectivity. To control for this, we tracked key postural markers using DeepLabCut[41,42] and combined them with in-task recorded instructed behaviours and task parameters to predict neural activity. We found that there was a significant and unique contribution for all instructed and uninstructed behavioural variables. Notably, however, there was also a unique contribution of the category component that could not be accounted for by any of the instructed or uninstructed behavioural parameters (Extended Data Fig. 10k–o, Supplementary Video 1). We therefore conclude that the mPFC contains a sparse but distinct set of neurons that represent learned categories irrespective of associated motor behaviours and reward.

Discussion

Using a paradigm to study learning of rule-based categories in mice, we could follow neuronal populations in the mPFC throughout the entire learning process, from naive to expert mice. We found two distinct groups of cells developing a representation of learned categories with different learning-related dynamics. The NoGo category representation emerged gradually, was rule-specific and was not strongly modulated by additional task parameters, in contrast to the Go representation. In addition, we observed that selectivity for the Go category increased further in the fast rule-switch phase compared to the slow, initial learning phase. This difference could be a consequence of Go category-selective neurons belonging to intrinsically different representations of choice, reward and categories. By experimentally decoupling these, we confirmed that many category-selective neurons were actually mixed-selective, which could benefit the representation of task-relevant information[38]. However, the experiment also revealed uniquely category-selective mPFC neurons, for both learned categories. In line with previous studies[3,19,32,43], we found that the mPFC initially contains a conjunctive stimulus and choice representation. This representation flexibly followed the novel Go category when a mouse learned the second rule. In parallel, a slowly learning group of Go category-selective cells emerged for each rule, following a time course similar to the NoGo category representation.

This mouse model of rule-based category learning opens up possibilities to causally investigate neuronal interactions across several cortical and subcortical circuits. Many brain areas, such as posterior parietal cortex[4,44], sensory areas (P.M.G., S.R., T.B. and M.H., manuscript submitted)[44,45] and striatum[3], contribute to multiple aspects of category learning and categorization behaviour. Several circuit models on areal interactions have been put forward[43,46]. One model of particular interest proposed that slow-learning PFC circuits acquire category selectivity using rapidly learned stimulus-specific activity originating in the striatum as a teaching signal[3,46]. Within this framework, the mPFC could compute the rule-dependent NoGo category representation from the fast-arising activity of conjunctive Go/choice-selective neurons mediated by local inhibitory circuits. Rule-based category learning in mice allows for testing of specific predictions of such circuit models for prefrontal cortex function by observing initially naive mice throughout the learning process. In particular, the possibility to investigate and observe neuronal responses in the mPFC during category learning in mice opens a window to study the neural circuitry that underlies categorization and storage of semantic memories[47] also in this species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03452-z.

1. Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract rules. *Nature* **411**, 953–956 (2001).
2. Freedman, D. J. & Assad, J. A. Experience-dependent representation of visual categories in parietal cortex. *Nature* **443**, 85–88 (2006).
3. Antzoulatos, E. G. & Miller, E. K. Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron* **71**, 243–249 (2011).

4. Brincat, S. L., Siegel, M., von Nicolai, C. & Miller, E. K. Gradual progression from sensory to task-related processing in cerebral cortex. *Proc. Natl Acad. Sci. USA* **115**, E7202–E7211 (2018).
5. Goodwin, S. J., Blackman, R. K., Sakellaridi, S. & Chafee, M. V. Executive control over cognition: stronger and earlier rule-based modulation of spatial category signals in prefrontal cortex relative to parietal cortex. *J. Neurosci*. **32**, 3499–3515 (2012).
6. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).
7. Roy, J. E., Riesenhuber, M., Poggio, T. & Miller, E. K. Prefrontal cortex activity during flexible categorization. *J. Neurosci*. **30**, 8519–8528 (2010).
8. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci*. **2**, 820–829 (2001).
9. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci*. **24**, 167–202 (2001).
10. Smith, J. D. et al. Implicit and explicit categorization: a tale of four species. *Neurosci. Biobehav. Rev*. **36**, 2355–2369 (2012).
11. Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J. & Ashby, F. G. Implicit and explicit category learning by macaques (*Macaca mulatta*) and humans (*Homo sapiens*). *J. Exp. Psychol. Anim. Behav. Process*. **36**, 54–65 (2010).
12. Shepard, R. N. & Chang, J.-J. Stimulus generalization in the learning of classifications. *J. Exp. Psychol*. **65**, 94–102 (1963).
13. Ashby, F. G. & Maddox, W. T. Human category learning. *Annu. Rev. Psychol*. **56**, 149–178 (2005).
14. Ashby, F. G. & Spiering, B. J. The neurobiology of category learning. *Behav. Cogn. Neurosci. Rev*. **3**, 101–113 (2004).
15. Smith, J. D. et al. Pigeons' categorization may be exclusively nonanalytic. *Psychon. Bull. Rev*. **18**, 414–421 (2011).
16. Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A. & Freeman, J. H. Selective attention in rat visual category learning. *Learn. Mem*. **26**, 84–92 (2019).
17. Jiang, X. et al. Categorization training results in shape- and category-selective human neural plasticity. *Neuron* **53**, 891–903 (2007).
18. Ragozzino, M. E., Detrick, S. & Kesner, R. P. Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. *J. Neurosci*. **19**, 4585–4594 (1999).
19. Rikhye, R. V., Gilra, A. & Halassa, M. M. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci*. **21**, 1753–1763 (2018).
20. de Bruin, J. P., Sànchez-Santed, F., Heinsbroek, R. P., Donker, A. & Postmes, P. A behavioural analysis of rats with damage to the medial prefrontal cortex using the Morris water maze: evidence for behavioural flexibility, but not for impaired spatial navigation. *Brain Res*. **652**, 323–333 (1994).
21. Mansouri, F. A., Matsumoto, K. & Tanaka, K. Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog. *J. Neurosci*. **26**, 2745–2756 (2006).
22. Brigman, J. L. & Rothblat, L. A. Stimulus specific deficit on visual reversal learning after lesions of medial prefrontal cortex in the mouse. *Behav. Brain Res*. **187**, 405–410 (2008).
23. Uylings, H. & van Eden, C. G. Chapter 3 Qualitative and quantitative comparison of the prefrontal cortex in rat and in primates, including humans. *Prog. Brain Res*. **85**, 31–62 (1991).
24. Chang, J.-Y., Chen, L., Luo, F., Shi, L.-H. & Woodward, D. J. Neuronal responses in the frontal cortico-basal ganglia system during delayed matching-to-sample task: ensemble recording in freely moving rats. *Exp. Brain Res*. **142**, 67–80 (2002).
25. Kesner, R. P. Subregional analysis of mnemonic functions of the prefrontal cortex in the rat. *Psychobiology (Austin Tex.)* **28**, 219–228 (2000).
26. Seamans, J. K., Lapish, C. C. & Durstewitz, D. Comparing the prefrontal cortex of rats and primates: insights from electrophysiology. *Neurotox. Res*. **14**, 249–262 (2008).
27. Carlén, M. What constitutes the prefrontal cortex? *Science* **358**, 478–482 (2017).
28. Laubach, M., Amarante, L. M., Swanson, K. & White, S. R. What, if anything, is rodent prefrontal cortex? *eNeuro* **5**, ENEURO.0315-18.2018 (2018).
29. Low, R. J., Gu, Y. & Tank, D. W. Cellular resolution optical access to brain regions in fissures: imaging medial prefrontal cortex and grid cells in entorhinal cortex. *Proc. Natl Acad. Sci. USA* **111**, 18739–18744 (2014).
30. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *J. Neurophysiol*. **88**, 929–941 (2002).
31. Fitzgerald, J. K. et al. Biased associative representations in parietal cortex. *Neuron* **77**, 180–191 (2013).
32. Pinto, L. & Dan, Y. Cell-type-specific activity in prefrontal cortex during goal-directed behavior. *Neuron* **87**, 437–450 (2015).
33. Otis, J. M. et al. Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. *Nature* **543**, 103–107 (2017).
34. Huda, R. et al. Distinct prefrontal top-down circuits differentially modulate sensorimotor behavior. *Nat. Commun*. **11**, 6007 (2020).
35. Li, B., Nguyen, T. P., Ma, C. & Dan, Y. Inhibition of impulsive action by projection-defined prefrontal pyramidal neurons. *Proc. Natl Acad. Sci. USA* **117**, 17278–17287 (2020).
36. Grunfeld, I. S. & Likhtik, E. Mixed selectivity encoding and action selection in the prefrontal cortex during threat assessment. *Curr. Opin. Neurobiol*. **49**, 108–115 (2018).
37. Karlsson, M. P., Tervo, D. G. & Karpova, A. Y. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135–139 (2012).
38. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
39. Freedman, D. J. & Assad, J. A. Neuronal mechanisms of visual categorization: an abstract view on decision making. *Annu. Rev. Neurosci*. **39**, 129–147 (2016).
40. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci*. **22**, 1677–1686 (2019).
41. Nath, T. et al. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protocols* **14**, 2152–2176 (2019).
42. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci*. **21**, 1281–1289 (2018).
43. Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U. & Waldron, E. M. A neuropsychological theory of multiple systems in category learning. *Psychol. Rev*. **105**, 442–481 (1998).
44. Zhong, L. et al. Causal contributions of parietal cortex to perceptual decision-making during stimulus categorization. *Nat. Neurosci*. **22**, 963–973 (2019).
45. Xin, Y. et al. Sensory-to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. *Neuron* **103**, 909–921.e6 (2019).
46. Villagrasa, F. et al. On the role of cortex-basal ganglia interactions for category learning: A neuro-computational approach. *J. Neurosci*. **38**, 9551–9562 (2018).
47. Tonegawa, S., Morrissey, M. D. & Kitamura, T. The role of engram cells in the systems consolidation of memory. *Nat. Rev. Neurosci*. **19**, 485–498 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data reporting

No statistical methods were used to predetermine sample size. Mice were randomly assigned to the categorization rule 'spatial frequency' or 'orientation'. The investigators were not blinded to allocation during experiments and outcome assessment.

Animals

All procedures were performed in accordance with the institutional guidelines of the Max Planck Society and the local government (Regierung von Oberbayern). Twenty female C57BL/6 mice (postnatal day (P) 63–P82 at the day of surgery) were housed in groups of four to six littermates in standard individually ventilated cages (IVC, Tecniplast GR900). Mice had access to a running wheel and other enrichment material such as a tunnel and a house. All mice were kept on an inverted 12 h light/12 h dark cycle with lights on at 22:00. Before and during the experiment, the mice had ad libitum access to standard chow (1310, Altromin Spezialfutter). Before the start of behavioural experiments, mice had ad libitum access to water. At the end of the experiments, mice were perfused with 4% paraformaldehyde (PFA) in PBS and their brains were post-fixed in 4% PFA in PBS at 4 °C.

Surgical procedures

Before surgery, a prism implant was prepared by attaching a 1.5 mm × 1.5 mm prism (aluminium coating on the long side, MPCH-1.5, IMM photonics) to a 0.13 mm thick, 3 mm diameter glass coverslip (41001103, Glaswarenfabrik Karl Hecht) using UV-curing optical glue (Norland optical adhesive 71, Norland Products) and was left to fully cure at room temperature for a minimum of 24 h. Mice were anaesthetized with a mixture of fentanyl, midazolam and medetomidine in saline (0.05 mg kg^{-1}, 5 mg kg^{-1} and 0.5 mg kg^{-1} respectively, injected intraperitoneally). As soon as sufficient depth of anaesthesia was confirmed by absence of the pedal reflex, carprofen in saline (5 mg kg^{-1}, injected subcutaneously) was administered for general analgesia. The eyes were covered with ophthalmic ointment (IsoptoMax/Bepanthen) and lidocaine (Aspen Pharma) was applied on and underneath the scalp for topical analgesia. The scull was exposed, dried and subsequently scraped with a scalpel to improve adherence of the head plate. The scalp surrounding the exposed area was adhered to the skull using Histoacryl (B. Braun Surgical). A custom-designed head plate was centred at ML 0 mm, approximately 3 mm posterior to bregma, attached with cyanoacrylate glue (Ultra Gel Matic, Pattex) and secured with dental acrylic (Paladur). A 3 mm diameter craniotomy, centred at anterior–posterior (AP) 1.9 mm, medial–lateral (ML) 0 mm, was performed using a dental drill. The hemisphere for prism insertion was selected based on the pattern of bridging veins. Before inserting the prism, two injections (50 nl min^{-1}) of 200–250 nl of virus solution (AAV2/1.hSyn.mRuby2.GSG.P2A.GCaMP6m.WPRE.SV40, titre: 1.02 × 10^{13} genome copies (GC) ml^{-1}, Plasmid catalogue 51473, Addgene) were targeted at the medial prefrontal cortex opposite to the prism implant, coordinates: AP 1.4 mm to AP 2.8 mm, ML 0.25 mm, dorsal–ventral (DV) 2.3 mm (Nanoject, Neurostar). The left hemisphere was injected in 11 mice, and the right hemisphere in 9 mice. Subsequently, a durotomy was performed using microscissors (15070-08, Fine Science Tools) over the contralateral hemisphere, next to the medial sinus. The prism implant was inserted, gently pushing the medial sinus aside until the target cortical region became visible through the prism (for a detailed description, see ref. [29]). The coverslip was attached to the surrounding skull using cyanoacrylate glue and dental acrylic. After surgery, the anaesthesia was antagonized with a mixture of naloxone, flumazenil and atipamezole in saline (1.2 mg kg, 0.5 mg kg^{-1} and 2.5 mg kg^{-1} respectively, injected subcutaneously) and the mice were placed under a heat lamp for recovery. Post-operative analgesia was provided for two subsequent days with carprofen (5 mg kg^{-1}, injected subcutaneously).

Visual stimuli

Stimuli for behavioural training were presented in the centre of a gamma corrected LCD monitor (Dell P2414H; resolution: 1,920 by 1,080 pixels; width: 52.8 cm; height: 29.6 cm; maximum luminance: 182.3 Cd m^{-2}). The centre of the monitor was positioned at about 0° azimuth and 0° elevation at a distance of 18 cm, facing the mouse straight on. The stimuli were 36 different sinusoidal gratings, each with a specific orientation and spatial frequency combination, shown in full contrast on a grey background (see Extended Data Fig. 1 for schematic of stimuli and task stages). Orientations ranged from 0° to 90°, the spatial frequencies from 0.023 cycles per degrees (cyc/°) to 0.25 cyc/° (orientations: [0, 15, 30, 60, 75, 90] °, spatial frequencies: [0.023, 0.027, 0.033, 0.06, 0.1, 0.25] cyc/°). The stimulus size was 45 retinal degrees in diameter, including an annulus of 4 degrees blending into the equiluminant grey background. The gratings drifted with a temporal frequency of 1.5 cycles per s.

In a subset of experiments ($n = 3$ mice), a dense stimulus space was presented, consisting of 49 stimuli ranging from 15° to 75° in orientation and from 0.027 cyc/° to 0.1 cyc/° in spatial frequency (orientations: [15, 30, 37.5, 45, 52.5, 60, 75]°, spatial frequencies: [0.027, 0.033, 0.036, 0.043, 0.052, 0.06, 0.1] cyc/°). Stimuli on the category boundary (either having an orientation of 45° or a spatial frequency of 0.043 cyc/°) were assigned to both categories, hence rewarded in 50% of trials.

All stimuli were created and presented using the Psychophysics Toolbox extensions of MATLAB[48–50].

Behaviour

Behavioural experiments started seven days after surgery. The water restriction regime and the behavioural apparatus were previously described[51]. In short, mice were restricted to 85% of their initial weight on the starting date by individually adjusting the daily water ration. First, mice were accustomed to the experimenter and head fixation in the setup by daily handling sessions lasting 10 min. During these sessions, the water ration was offered in a handheld syringe. The remainder was supplemented in an individual drinking cage after a delay of approximately 30 min. After four to seven days of handling, mice were pre-trained to lick for reward, while being head-fixed on the spherical treadmill[52–54] in absence of visual stimulation. Whenever a mouse ceased to run (velocity below 1 cm s^{-1}) and made a lick on the spout, a water reward (drop size 8 µl) was delivered via the spout. A baseline imaging time point (T1) was acquired once the mice consumed more than 50 drops per session (35 to 45 min) on two consecutive days (requiring about three days of pre-training).

Subsequently, daily sessions of visual discrimination training for two initial stimuli started. Each mouse was randomly assigned to one of two groups. One group was first trained on the orientation rule, then on the spatial frequency rule. For the other group, the sequence of the rules was reversed (Extended Data Fig. 1). Each rule defined a Go category and a NoGo category, separated by a boundary at either 45° (orientation rule) or at 0.043 cyc/° (spatial frequency rule). Trials started with an inter-trial interval of 5 s. After that, the mouse could initiate stimulus presentation by halting and refraining from licking for a minimum of 0.5 s. A single stimulus was subsequently shown for 1.3 ± 0.2 s. At any time during stimulus presentation, the mouse could make a lick to indicate a Go choice. Trials with a Go choice in response to a Go category stimulus triggered a water reward and were classified as hits; trials in which the mice failed to lick during Go category stimulus presentation were considered misses. Correct withholding of a lick to a NoGo category stimulus was classified as a correct rejection, and did not result in a water reward. A lick during a NoGo category stimulus counted as a false alarm. Initially, false alarms only led to the termination of the current trial; later during training, false alarms were followed by a time-out of 5–7 s showing a time-out stimulus (a narrow, horizontal, black bar). Time-outs were included to reduce a Go bias that mice typically showed. The second imaging session (T2) was carried out after a mouse performed at more than 66% correct Go choices in a given session (requiring 11 to 40 sessions).

For the next training stage (leading up to imaging session T3) further stimuli were added (Extended Data Fig. 1a), such that both the Go category and the NoGo category consisted of three stimuli differing in the feature either irrelevant to the category rule (T3a, $n = 6$ mice), or relevant to the category rule (T3b, $n = 5$ mice). Whenever a mouse's performance exceeded 66% correct Go choices in one session, we proceeded to the next training (and imaging) stage; 6 stimuli per category, 9 stimuli per category (imaging session T4), and finally 18 stimuli per category (imaging session T5), the latter serving as a crucial test for generalization behaviour.

Rule-switch: After successful learning of rule 1, mice ($n = 11$) were retrained using the previously irrelevant dimension. This stage, known as rule-switch training, started with two exemplar stimuli for the new rule, and then proceeded with the same steps as for rule 1 and ended with another generalization test of rule 2 (18 stimuli per category, imaging session T8).

Task change: After successful learning of rule 1 (T5), the categorization performance of mice ($n = 9$) was tested with a different operant response, in a left/right choice task. For this session, the behavioural setup was slightly modified to create a left/right choice task. Instead of one lick spout centred in front of the mouse, the mouse was now presented with two lick-spouts, one offset to the left and one offset to the right. Stimuli of the previous Go category were assigned a new GoRight response (rewarded after a lick on the right lick spout). Stimuli of the previous NoGo category were assigned a new GoLeft response (rewarded after a lick on the left lick spout). The original stimulus to category assignment—that is, the categorization rule—remained the same throughout the task change. Before the first stimulus presentation, ten drops were manually given on each lick spout to motivate the mice to lick on both sides.

Throughout training, stimuli from the Go category and the NoGo category were presented in a pseudorandomized fashion, showing not more than three stimuli of the same category in a row. The behavioural training program was a custom written MATLAB routine (Mathworks).

Imaging

Two-photon imaging[55] through the implanted prism was performed at 5–8 time points in each mouse throughout the training paradigm (T3 was omitted in two mice; for detailed timing of the imaging sessions see Extended Data Fig. 1a). In some mice ($n = 5$) we followed two regions in the same mouse; in these cases, two imaging sessions were acquired on consecutive days during the same training stage. Imaging was done using a custom-built two-photon laser-scanning microscope (resonant scanning system) and a Mai Tai eHP Ti:Sapphire laser (Spectra-Physics) tuned to a wavelength of 940 nm. Images were acquired with a sampling frequency of 10 Hz and 750 × 800 pixels per frame. The mice in the task change experiment were imaged using a customized commercially available two-photon laser-scanning microscope (Thorlabs; same laser specifications as described above), operated with Scanimage 4[56]. In these experiments, images were acquired at 30 Hz and 512 × 512 pixels per frame. The average laser power under the objective ranged from 50 to 80 mW. Note that the laser power was higher than for imaging through a conventional cranial window due to a substantial power loss over the prism[29]. We used a 16×, 0.8 NA, water immersion objective (Nikon) and diluted ultrasound gel (Dahlhausen) on top of the implant as immersion medium. Two photomultiplier tubes detected the red fluorescence signal of the structural protein mRuby2 (570–690 nm) and the green fluorescence signal of GCaMP6m (500–550 nm)[57]. During imaging, the monitor used for stimulus presentation was shuttered to minimize light contamination[58]. The imaging data were acquired using custom LABVIEW software (National Instruments; software modified from the colibri package by C. Seebacher) and the synchronization of imaging data with behavioural readout and stimulus presentation was done using DAQ cards (National Instruments).

Tracking of postural markers

In two-photon imaging sessions of a subset of experiments, the mouse was video-tracked using infrared cameras (The Imaging Source Europe). Two cameras were aimed at the eyes, and a third camera was positioned at a slight angle behind the mouse, in order to record body movements in-task. The eyes of the mouse were back-lit by the infrared two-photon imaging laser and the body was illuminated using an infrared light source (740 nm; Thorlabs). Key eye and body features (see Extended Data Fig. 10) were manually defined and automatically annotated using DeepLabCut[41,42]. From the x and y coordinates of these features, we calculated three eye parameters and four postural parameters (pupil diameter, eye position, eyelid opening, front paw angle, hind paw angle of the left hind paw, body elongation/rotation, tail angle; see Extended Data Fig. 10). Supplementary Video 1 shows both eye and body cameras of an example mouse.

Data analysis

The analysis of behaviour and imaging data was performed using custom written MATLAB routines.

Behavioural data

Behavioural performance is shown as the sensitivity index, d'. For every training session, d' was calculated as the difference between the z-scored hit rate and the z-scored false alarm rate. The hit rate was defined as the number of correct category 2 trials divided by the total number of category 2 trials per session. Similarly, the false alarm rate was calculated as the number of incorrect category 1 trials divided by the total number of category 1 trials. In case a mouse performed two training sessions at time points T1, T3, T4, T5, T7 and T8, because two regions were imaged, the displayed value in the learning curve is the average across the two imaging sessions.

The fraction of correct Go choices was calculated as the number of hit trials divided by the number of all trials in which the mouse made a Go choice (the sum of 'hits' and 'false alarms'). The number of days until a mouse reached performance criterion was the amount of daily sessions until the fraction of correct Go choices exceeded 0.66. Pre-training sessions without visual stimulation were not included in this quantification.

To investigate categorization behaviour across the entire stimulus space, we calculated the 'fraction chosen': The number of Go choices in response to a specific stimulus divided by the total number of presentations for that stimulus (see example in Fig. 1d; for all mice see Extended Data Fig. 2). Finally, we constructed psychometric curves showing the effect of each feature (that is, rule-relevant versus rule-irrelevant) on the behaviour of the mice (Fig. 1j). For that, the stimulus-specific 'fraction chosen' values were averaged along the irrelevant or the relevant feature dimension, respectively (see Fig. 1i).

To estimate learning rates, each individual learning curve was fitted with a sigmoid function:

$$y(x) = p1 + \frac{p2}{1 + e^{p3(x-p4)}}$$

in which $p1$ determines the minimum of the sigmoid curve (for curve fitting fixed to 0), $p2$ the maximum, $p3$ the slope and $p4$ the inflection point. The parameter defining the minimum was fixed at a d' of 0. Learning curves for rule 1 and rule 2 were fitted independently. Goodness of fit was determined as the root-mean-square error between the learning curve and the fitted curve.

Imaging data processing

The imaging data were first preprocessed by performing dark-current subtraction (using the average signal intensity during a laser-off period) and line shift correction. Rigid xy image displacement was first calculated on the structural red fluorescence channel using the cross correlation of the 2D Fourier transform of the images[59], and subsequently corrected on both channels. For each imaging session, cells were manually segmented using the average image of the red fluorescence channel across the entire session. The cell identity was then manually matched across all imaging time points and only cells that could be identified in every session from T1 to T8 or T5 to left/right were included in the analysis. This criterion excluded one mouse (M06) from all further analyses, because of lost optical access at T8. The average green fluorescence signal was extracted for each cell and then corrected for neuropil contamination by subtracting the signal of 30 μm surrounding each cell multiplied by 0.7 and adding the median multiplied by 0.7 (refs. [57,60]). From this fluorescence trace, we calculated $\Delta F/F$ as $(F - F_0)/F_0$ per frame. F_0 was defined as the 25th percentile of the fluorescence trace in a sliding window of 60 s. From this trace, we inferred the spiking activity of each cell using the constrained foopsi algorithm[61–63]. The inferred spike rate during the stimulus presentation period was used for all further calculations and in all figure panels, except for the HLS maps and the left panels of Fig. 2d, e, where we display the $\Delta F/F$ trace for comparison.

To display lick-triggered neuronal activity (Extended Data Fig. 8), we averaged the inferred spike rate centred on the onset of the mouse's lick-bouts. A lick-bout was defined as a sequence of licks, in which the interval between every two consecutive licks did not exceed 500 ms. Thus, a lick was part of a lick-bout if it happened within 500 ms after the previous lick. The onset of each lick-bout was the time of the first lick in the lick-bout.

Category-tuning index

For every cell, we calculated the CTI as previously described[30]. In short, we quantified the mean inferred spike rate during stimulus presentation for every stimulus. Next, we calculated the mean difference in inferred rate between stimuli of the same category (within), subtracted it from the mean difference between stimuli belonging to the two different categories (across) and normalized by the sum (across + within). This calculation results in an index ranging from −1 to 1, with category-unselective cells showing CTIs close to and below 0 and an ideal category-selective cell having an index of 1. Category-selective cells were defined as cells with a CTI value larger than 0.1. This threshold was chosen based on the distribution of CTIs in the naive population (T1), where individual cells rarely crossed this value. As a control, we used other thresholds (0.07, 0.15 and 0.20) and found no qualitative difference in the results other than that the fraction of category-selective cells scaled.

The fraction of category-selective cells was calculated as the number of neurons above threshold per imaging region, divided by the total number of chronically recorded neurons in that imaging region. Category-selective cells, determined by their CTI at time points T5 and T8, were divided in a Go category-selective and a NoGo category-selective group; neurons with higher average activity in Go category trials than in NoGo category trials were grouped as Go category-selective cells and conversely, cells with a higher average activity in NoGo category trials were labelled as NoGo category-selective. The overlap between the Go and NoGo category-selective groups was calculated between T5 and T8. The expected range of overlap assuming random independent sampling was calculated from the data, but with shuffled neuron identities (using the 95% percentile of the shuffled distribution). For time points at which not all stimuli were presented (T2, T3, T4, T6 and T7), we approximated category-tuning from the average responses to Go category trials and NoGo category trials.

Bayesian decoding

We decoded category identity from trial-by-trial activity patterns of a single neuron up to groups of ten neurons using Bayes theorem:

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)}$$

in which $p(r|c)$ is the probability of a single trial response r when observed in either category 1 or 2 trials (calculated from an exponential distribution), $p(c)$ as the prior probability of observing each category, and $p(r)$ as the probability of observing the response. To cross-validate decoding performance, trials were first split into a training and test set (70% and 30%, respectively). The trial-averaged inferred spike rates followed an exponential distribution, which we estimated for each category individually (using the training set). Then, for each trial in the test set, we calculated the probability that the neuronal response came from those distributions. The distribution that gave the higher probability was determined as the decoder's prediction. Decoder performance was calculated as the fraction of correctly predicted trials. As a control, decoding performance was also calculated after shuffling category identities across trials.

Selectivity time course

Average selectivity of individual neurons was calculated as the mean difference between responses to all Go category stimuli and all NoGo category stimuli, at every imaging time point (T1–T8). For linear regression, we defined three characteristic selectivity time courses (shown in Extended Data Fig. 7), resembling acquired selectivity for reward/choice, categorization rule 1 and categorization rule 2. Within each of these time courses, maximum selectivity was assigned the value 1 and no selectivity the value 0. The characteristic time courses were used as predictors in a model fitting the development of selectivity of individual neurons over time.

Generalized linear models to assess the influence of individual task parameters

We performed multilinear regression on neurons that were identified in all imaging time points of the rule-switch experiment. The regression model predicted the trial-wise mean spike rate of each cell during the stimulus presentation periods at imaging time point T5. Categorical predictors were: Category identity of the presented stimulus (0: category 1, 1: category 2), choice of the mouse (0: NoGo, 1: Go), and reward (0: no reward, 1: reward). The average running speed during the trial was modelled as a continuous predictor. A positive predictor weight indicated that the activity of a neuron was increased in trials where the value of the predictor was higher. A negative predictor weight reflected an inverse relation between the predictor's value and the neuron's firing rate. We normalized the predictor weights for overall differences in response amplitudes, by dividing each weight by the sum of all absolute predictor weights (including the intercept).

Hierarchical clustering was performed on relative predictor weights of neurons, including only cells with an R^2 value larger than 0.05. The optimal number of clusters was calculated using gap statistic values, determined as the smallest cluster number k that fulfilled the criterion (here nine clusters):

$$\text{Gap}(k) \geq \text{Gapmax} - \text{s.e. (Gapmax)}$$

in which Gap(k) is the gap statistic for k clusters, Gapmax is the largest gap value, and s.e.(Gapmax) is the standard error corresponding to the largest gap value.

We obtained linkage and relative predictor weights of the clusters from the MATLAB clusterdata algorithm.

To probe the influence of operant motor behaviour in the task change experiment, we concatenated all trials of sessions T5 (generalization session, Go/NoGo task) and L/R (left/right choice task). A stepwise linear regression model predicted the trial-averaged inferred spike rate of all recorded neurons individually. The predictors were the following categorical variables: category identity of the stimulus (0: category 1; 1: category 2), Go response of the mouse (0: NoGo, 1: all forms of Go, that is, Go/GoRight/GoLeft), reward (0: no reward, 1: reward) and two predictors that were specific to a motor response in the left/right session: GoRight and GoLeft. We only considered significant predictor weights, determined from an *F*-statistic comparing a model with and without a predictor. Predictor weights were normalized by dividing each weight by the maximum of all predictor weights.

Linear regression assessing the influence of instructed and uninstructed behaviours

The trial-averaged inferred spike rate of all recorded neurons in session T5 of a subset of experiments was fitted using a linear model. Body and eye parameters describing uninstructed behaviours were included in the model as continuous predictors. In addition, we included three categorical task-relevant predictors: category identity of the presented stimulus, choice of the mouse, and reward. For each predictor, we determined its maximum predictive power (cvR^2) and its unique contribution (ΔR^2), similar to the approach previously described[40]. Maximum predictive power (cvR^2) was calculated as the predictive performance (R^2) of a model with all parameters shuffled, except for the parameter of interest. A parameter's unique contribution (ΔR^2) was quantified as the difference between the full model's R^2 and the R^2 of a model in which the parameter of interest was shuffled.

Stereotaxic coordinates of imaging regions

We determined the stereotaxic coordinates of the centres of all imaging regions (included in Fig. 2g, h) to place the imaged regions within a common reference frame (Mouse Brain Atlas)[64]. First, we cut 60-µm thick sagittal sections of both hemispheres using a freezing microtome. The AP coordinates outlining the full extent of the prism were identified from a section of the hemisphere into which the prism had been implanted (Extended Data Fig. 4). On the basis of this information, we calculated the exact AP coordinate of the centre of each imaging field of view. We calculated the dorso-ventral coordinate relative to the brain surface, which was aligned with the dorsal border of the prism. Finally, we determined the medio-lateral coordinate of the imaged field of view from the imaging depth of the field of view relative to the medial pia mater.

Statistical procedures

All data are presented as mean ± s.e.m. unless stated otherwise. Tests for normal distribution were carried out using the Kolmogorov–Smirnov test. Normally distributed data were tested using the two-tailed paired-samples *t*-test. Non-normally distributed data were tested using the two-tailed WMPSR test for paired samples, and the Kruskal–Wallis test for multiple, independent groups. A Bonferroni alpha correction was applied when multiple tests were done on the same data. Correlations were assessed using Pearson's correlation coefficient, if the data were normally distributed along both axes; otherwise, Spearman's correlation was applied.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data supporting the findings of this study are available on publication at https://gin.g-node.org/sreinert/Category-learning_mPFC. Source data are provided with this paper.

Code availability

The custom written MATLAB routines used for data collection and analysis are available upon reasonable request.

48. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
49. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–42 (1997).
50. Kleiner, D. et al. What's new in psychtoolbox-3? *Perception* **36**, 1–16 (2007).
51. Goltstein, P. M., Reinert, S., Glas, A., Bonhoeffer, T. & Hübener, M. Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice. *PLoS ONE* **13**, e0204066 (2018).
52. Hölscher, C., Schnee, A., Dahmen, H., Setia, L. & Mallot, H. A. Rats are able to navigate in virtual environments. *J. Exp. Biol.* **208**, 561–569 (2005).
53. Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L. & Tank, D. W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56**, 43–57 (2007).
54. Guo, Z. V. et al. Procedures for behavioral experiments in head-fixed mice. *PLoS ONE* **9**, e88678 (2014).
55. Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248**, 73–76 (1990).
56. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
57. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
58. Leinweber, M. et al. Two-photon calcium imaging in mice navigating a virtual reality environment. *J. Vis. Exp.* **84**, e50885 (2014).
59. Guizar-Sicairos, M., Thurman, S. T. & Fienup, J. R. Efficient subpixel image registration algorithms. *Opt. Lett.* **33**, 156–158 (2008).
60. Kerlin, A. M., Andermann, M. L., Berezovskii, V. K. & Reid, R. C. Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron* **67**, 858–871 (2010).
61. Vogelstein, J. T. et al. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104**, 3691–704 (2010).
62. Pnevmatikakis, E. A. et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89**, 285–299 (2016).
63. Giovannucci, A. et al. CaImAn an open source tool for scalable calcium imaging data analysis. *eLife* **8**, e38173 (2019).
64. Franklin, K. B. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* Vol. 3 (Academic, 2007).

Acknowledgements We thank C. Leibold for discussions contributing to the study design; V. Staiger, C. Huber, F. Voss and M. Sperling for technical assistance; A. Glas for help with the implant method and discussions; J. Bauer, D. Mearns and T. Mrsic-Flogel for discussions and comments on the manuscript; T. Rose for the viral construct; J. Kuhl for illustrations. This project was funded by the Max Planck Society and the Collaborative Research Center SFB870 (project numbers A07, A08) of the German Research Association (DFG) to T.B. and M.H. and the German Research Association (DFG) via the RTG 2175 'Perception in context and its neural basis' to M.H. and C. Leibold.

Author contributions S.R., M.H., T.B. and P.M.G. designed the study. S.R. conducted the experiments. S.R. and P.M.G. analysed the data. S.R., M.H., T.B. and P.M.G. wrote the manuscript.

Funding Open access funding provided by Max Planck Society.

Competing interests The authors declare no competing interests.

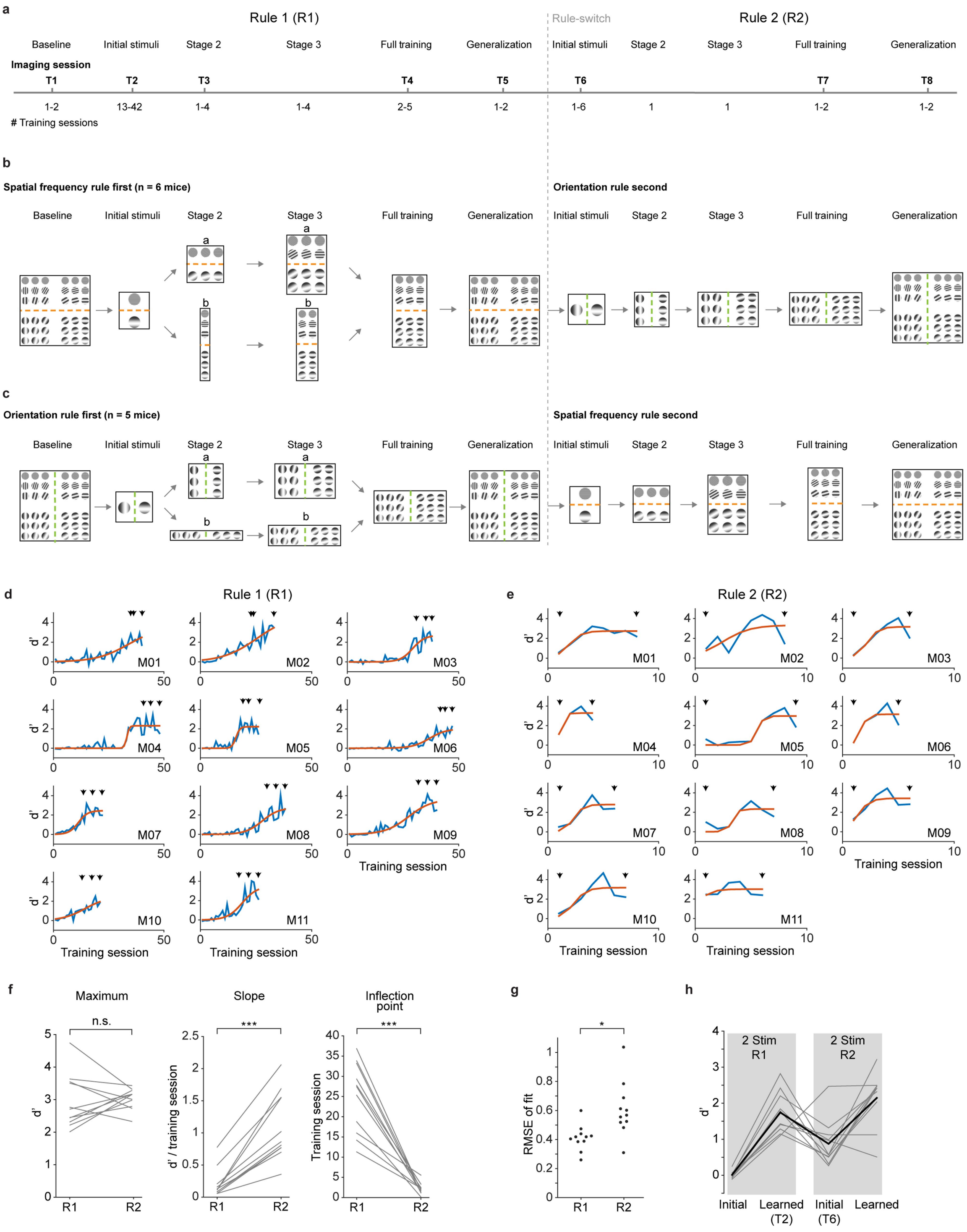
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03452-z.

Correspondence and requests for materials should be addressed to P.M.G.

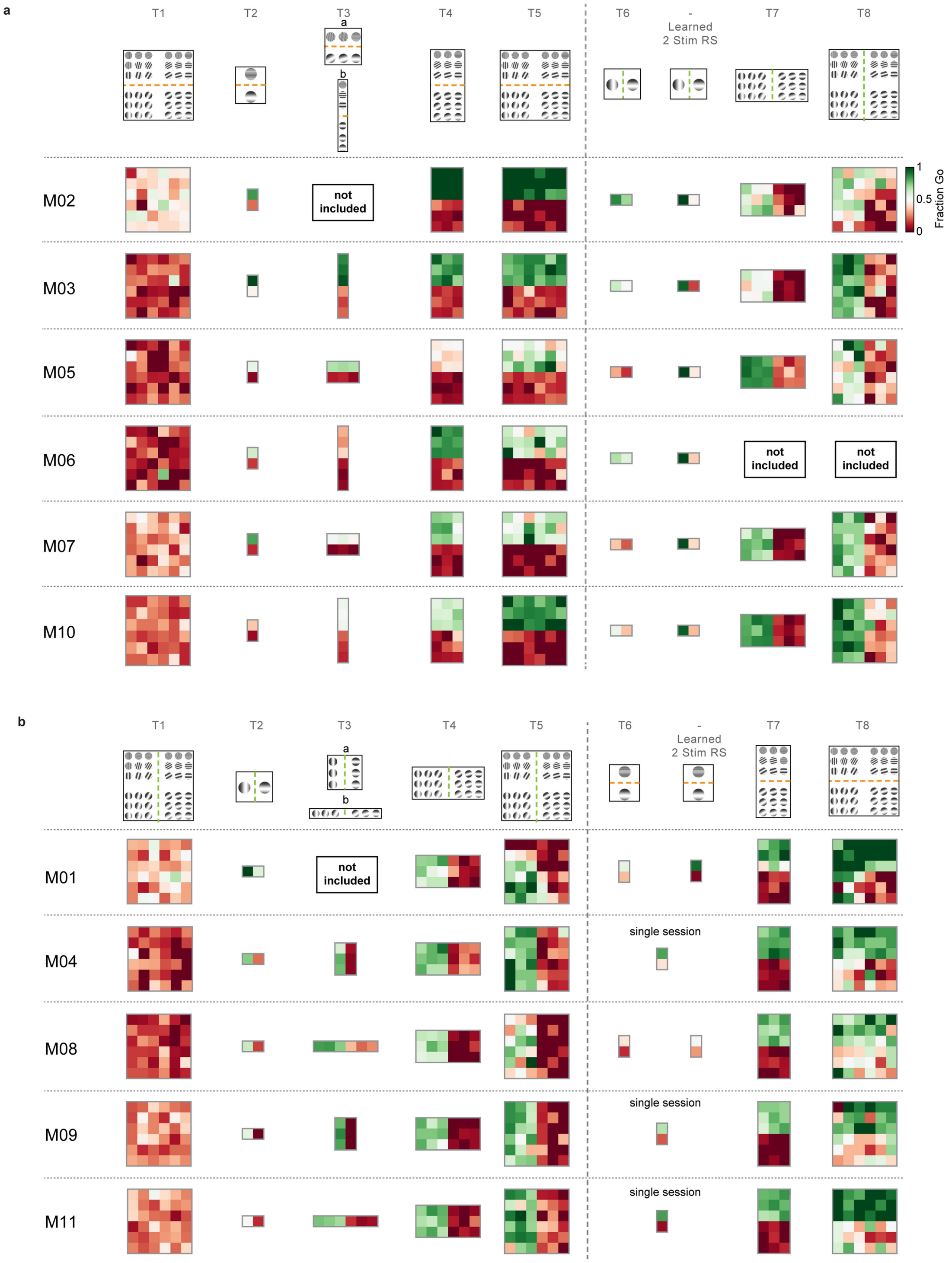
Peer review information *Nature* thanks Anne Churchland and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

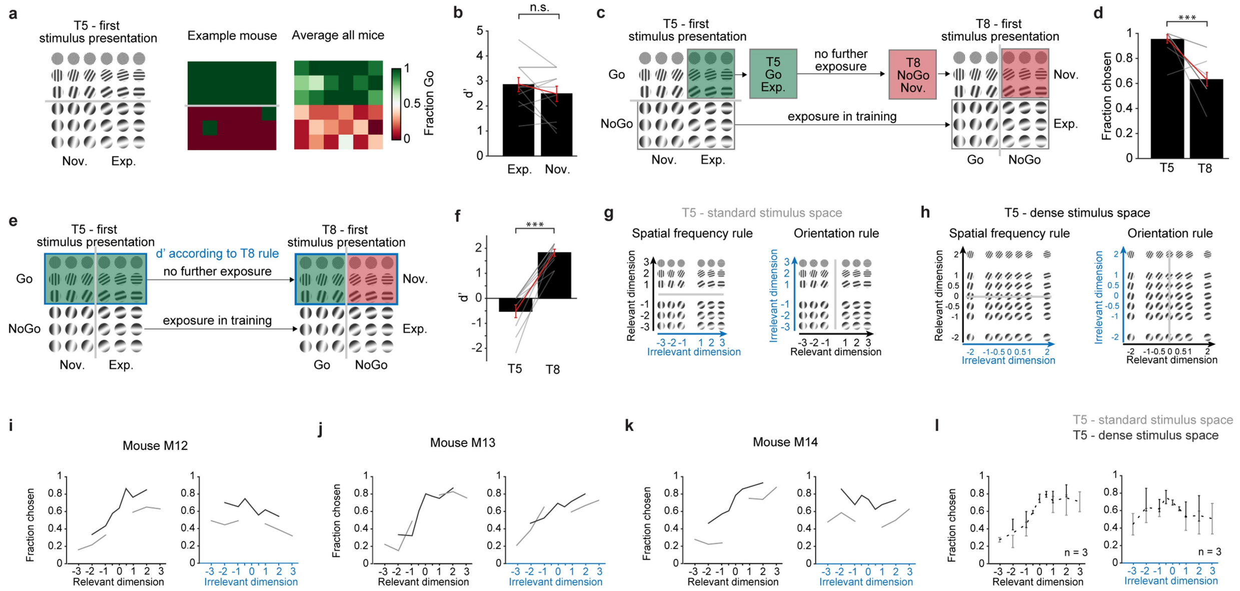


Extended Data Fig. 1 | See next page for caption.

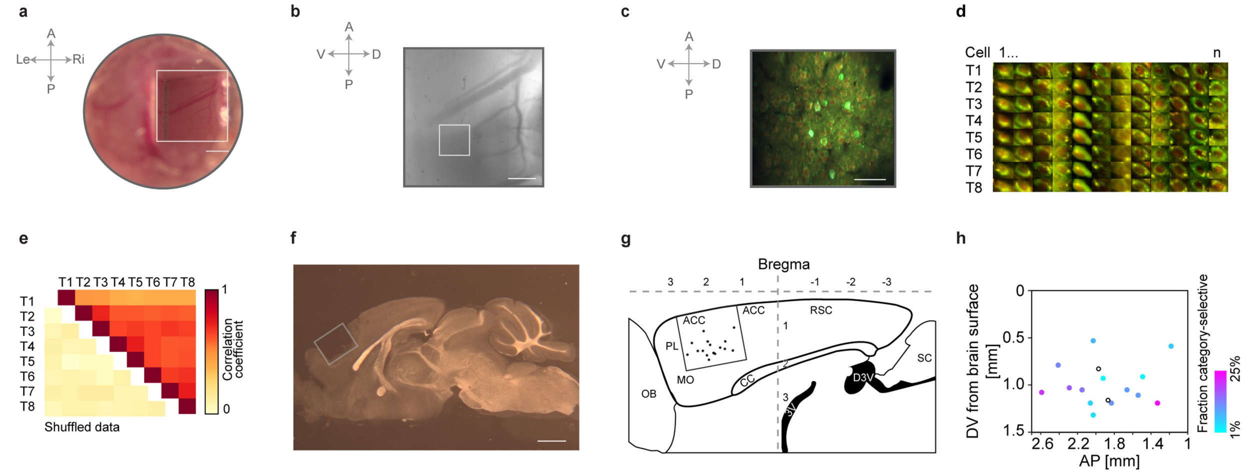
Extended Data Fig. 1 | Timeline of behavioural training, presented stimuli and learning performance of individual mice for both categorization rules. **a**, Timeline showing behavioural training stages, the number of training sessions that mice spent in each stage (min–max) and the imaging sessions (T1–T8). **b**, Stimuli used for category training, aligned to the stages shown in **a**. The scheme shows stimuli for mice that were trained on the spatial frequency rule first, and the orientation rule second. **c**, As in **b**, but for mice trained on the orientation rule first. **d**, Per mouse, the learning curve for training on rule 1. Blue curve denotes single session *d′*. Orange curve denotes sigmoid fit of *d′*. Arrows indicate imaging time points T2, T3 and T4. **e**, As in **d**, but for rule 2. Arrows indicate imaging time points T6 and T7. **f**, Parameters describing the fitted sigmoidal curves, comparing rules 1 and 2. Left, maximum. $P = 0.71$, two-tailed paired-samples *t*-test ($n = 11$). Middle, slope. $P = 1.9 \times 10^{-5}$, two-tailed paired-samples *t*-test ($n = 11$). Right, inflection point. $P = 9.3 \times 10^{-6}$, two-tailed paired-samples *t*-test ($n = 11$). **g**, Root-mean-square error (RMSE) of sigmoid fit. $P = 0.013$, two-tailed paired-samples *t*-test ($n = 11$). **h**, *d′* of all mice comparing naive and learned discrimination of the initial two stimuli for the first rule (left) and the second rule (right). Black line indicates the mean across all mice, grey lines represent data of individual mice.



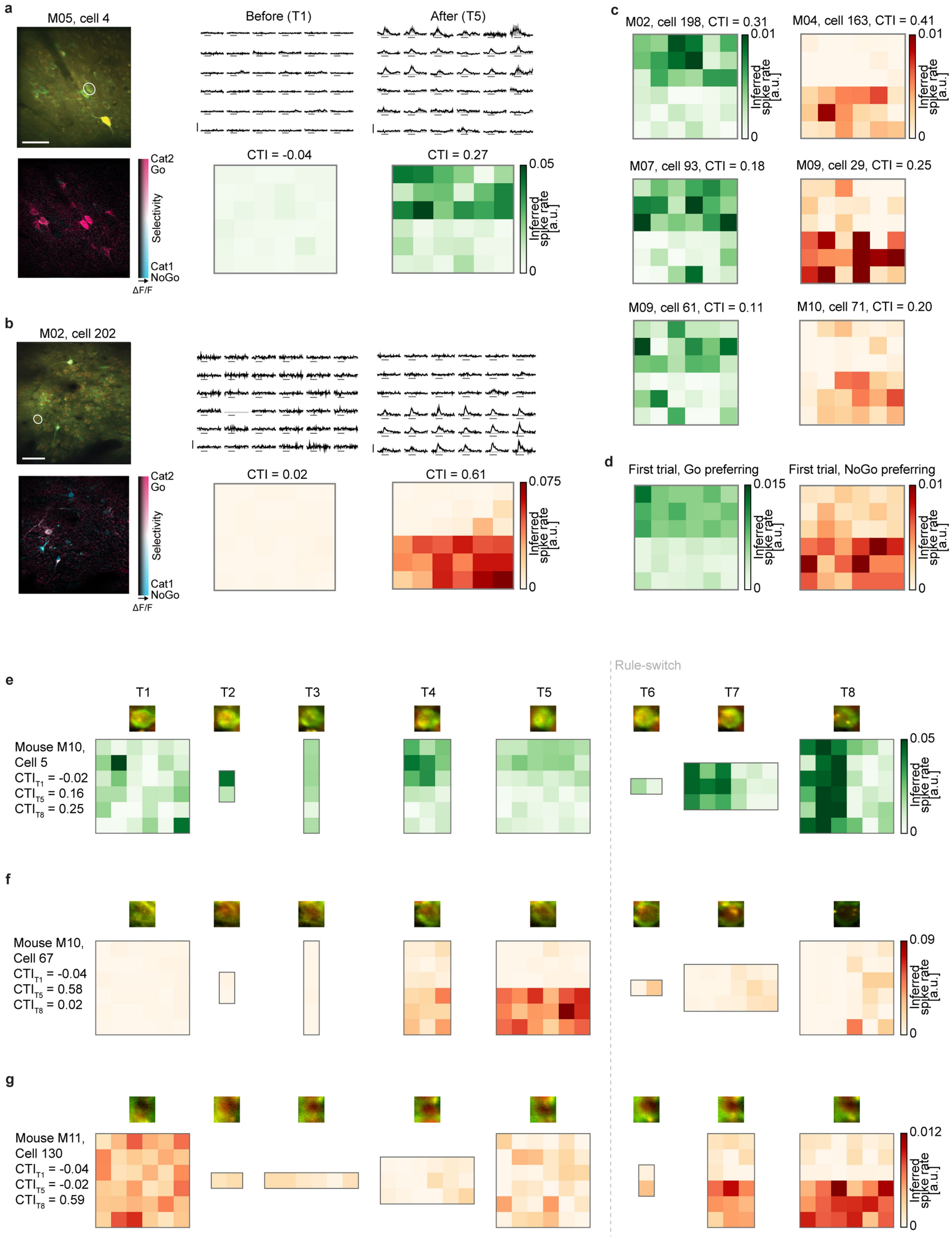
Extended Data Fig. 2 | Categorization, generalization and rule-switch performance for individual mice. Performance as the fraction of Go choices per stimulus, averaged over the imaging time points for each mouse individually. The time point 'Learned 2 stim RS' shows performance after the rule-switch was successfully learned. This time point was not an imaging session. Three mice learned the rule-switch during session T6 ('single session'). **a**, Mice first trained on the spatial frequency rule and then on the orientation rule (data of M03 is also shown in Fig. 1d, g). **b**, As in **a** for all mice trained initially on the orientation rule and then on the spatial frequency rule. Mouse M06: imaging sessions T7 and T8 were 'not included' owing to poor imaging quality; Mice M01 and M02: imaging session T3 was not recorded.



Extended Data Fig. 3 | Generalization of stimuli at their first presentation and categorization of stimuli close to, and at the category boundary. **a**, Left, schematic of stimulus space during the generalization session (T5). Middle, category choice for every stimulus on its first presentation for an example mouse, green: Go choice, red: NoGo choice. Right, category choice at the first stimulus presentation, averaged across mice ($n = 10$ mice). **b**, d' for experienced stimuli and novel stimuli separately, calculated using only the first presentation of each stimulus at the generalization session T5. $P = 0.19$, two-tailed paired-samples t-test ($n = 10$). Grey lines denote individual mice. Data are mean ± s.e.m. (across mice). **c**, Mice use the second rule to categorize stimuli that were only experienced during training on the first rule. Left, schematic showing category identity of stimuli at T5 (Go or NoGo) and whether they were experienced throughout category training on rule 1 (Exp) or novel (Nov). Middle, the highlighted quadrant (green) was part of the Go category; stimuli from this quadrant had been incrementally used throughout category learning up to T5. After T5, mice were trained on the second rule, using only stimuli in the bottom half of the category space (which corresponded to the NoGo category at T5). Right, in the second generalization session (T8, rule-switch generalization), mice were once more exposed to the full category space. Now, the same highlighted quadrant (red) required a NoGo response. However, so far these stimuli were extensively (and only) experienced as requiring a Go response. If mice showed a different fraction of Go choices in T5 and T8, it would reflect category generalization of rule 2, because the absence of experience with the stimuli in the highlighted quadrant prevented learning of a stimulus-response mapping. **d**, Fraction of Go category choices for the first presentation of stimuli highlighted in **c**, comparing T5 and T8. Data are mean ± s.e.m. (fraction chosen across all mice). Grey lines indicate data of individual mice, of which some overlap. $P = 0.002$, two-tailed paired-samples t-test ($n = 10$). **e**, Schematic as in **c**, indicating the stimuli and the rule for which d' was calculated in **f**. **f**, d' for the first presentation of stimuli highlighted in **c**, comparing T5 and T8. Grey lines denote individual mice. $P = 2.2 \times 10^{-5}$, two-tailed paired-samples t-test ($n = 10$). Data are mean ± s.e.m. (across mice). **g**, Schematic indicating the relevant and irrelevant stimulus dimension for the spatial frequency rule (left) and the orientation rule (right) at T5. **h**, As in **g**, for training with a dense stimulus space ($n = 3$ mice) to determine categorization behaviour closer to the category boundary. **i**–**k**, Psychometric curves for the three individual mice. The fraction chosen (fraction of Go choices) is shown along the relevant dimension (left) and the irrelevant dimension (right). Grey lines denote data from the T5 generalization session. Black lines denote data from the session with the dense stimulus space. **l**, As in **i**, showing the mean (± s.e.m.) across the three mice (shown in **i**, **j** and **k**) that were tested using the dense stimulus space.

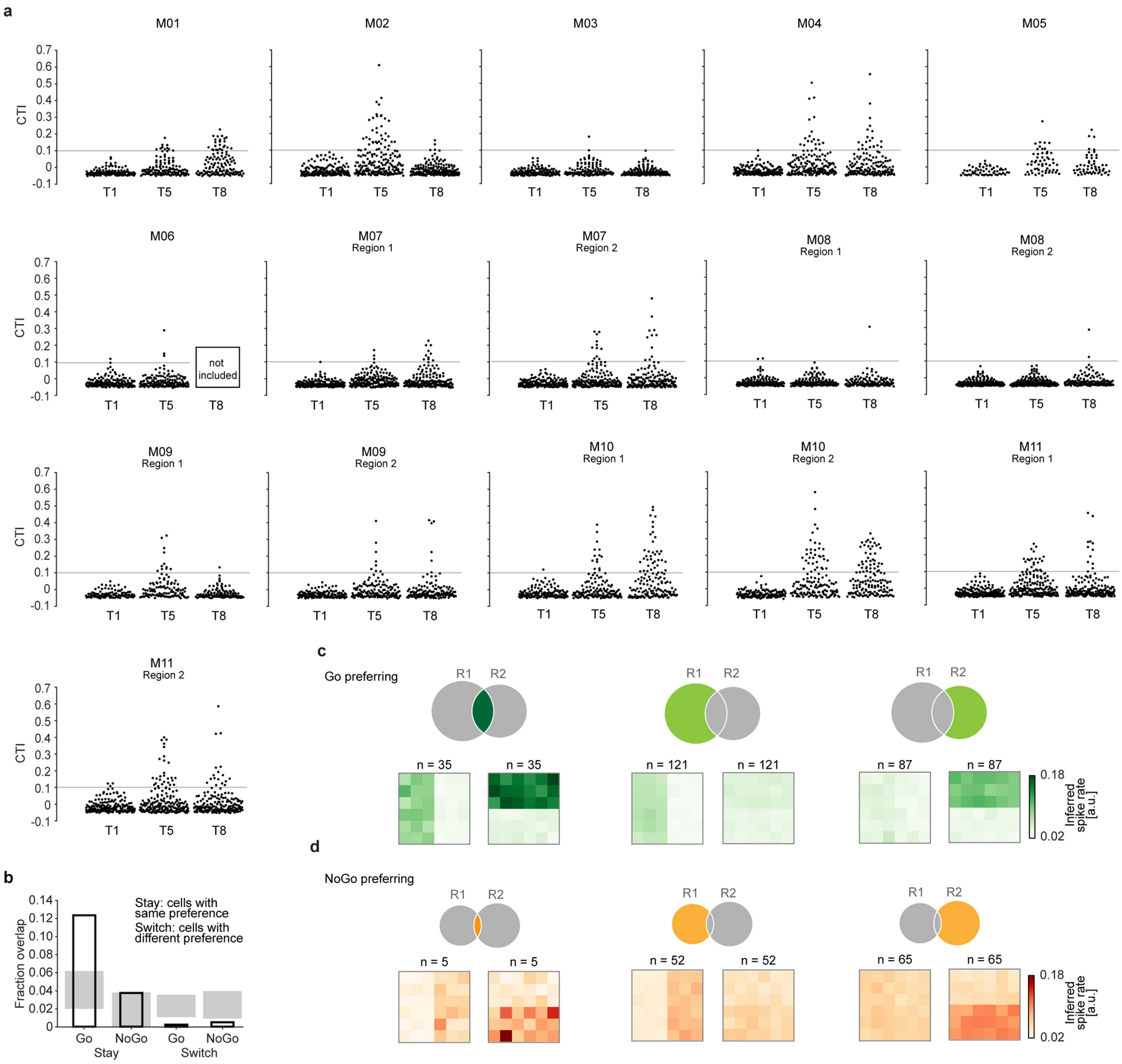


Extended Data Fig. 4 | Reconstruction of the location of imaging regions. **a**, Top down view onto the craniotomy of M07 with prism implant (white square denotes the prism outline). A, anterior; P, posterior; Le, left; Ri, right. Scale bar, 0.5 mm. **b**, View through the prism with the position of an imaging field (white box), D, dorsal; V, ventral. Scale bar, 0.3 mm **c**, The imaging field in **b**, visualized with a two-photon microscope (red: structural marker mRuby2; green: functional marker GCaMP6m; image is the average of all frames of session T1). Scale bar, 30 µm. **d**, Cropped images showing 12 example neurons across all imaging time points (T1–T8). **e**, The top triangle shows the correlation between cropped images of any two time points (average across all neurons). The bottom triangle shows the correlation after shuffling cell identities (control). **f**, Example sagittal brain section showing the position of the prism implant along the anterior-posterior axis. Scale bar, 1 mm. **g**, Schematic of cortical midline regions near the prism implant (ML 0.12), modified from Franklin & Paxinos[64], figure 102, with permission from Academic Press (Copyright 2007). 3V, third ventricle; ACC, anterior cingulate cortex; CC, corpus callosum; D3V, dorsal third ventricle; MO, medial orbital cortex; OB, olfactory bulb; PL, prelimbic cortex; RSC, retrosplenial cortex; SC, superior colliculus. The centres of all imaging regions included in Fig. 2g, h are indicated by black dots. **h**, Fraction of category-selective cells for each imaged field of view (included in Fig. 2g, h). The black, hollow circles are imaging regions without category-selective cells. There was no clear relationship between the location of the imaging regions within mPFC and the fraction of category-selective cells.

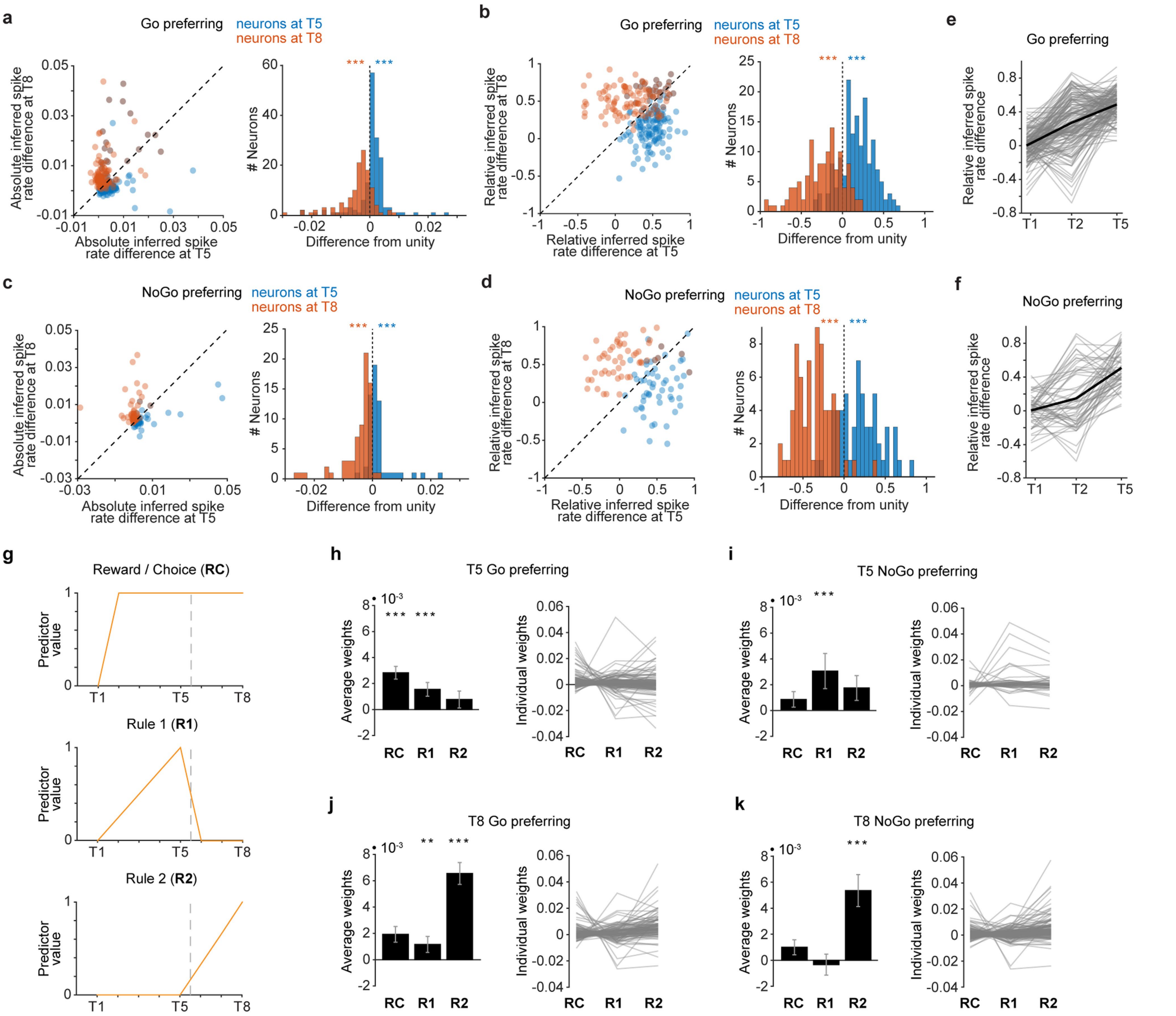


Extended Data Fig. 5 | See next page for caption.

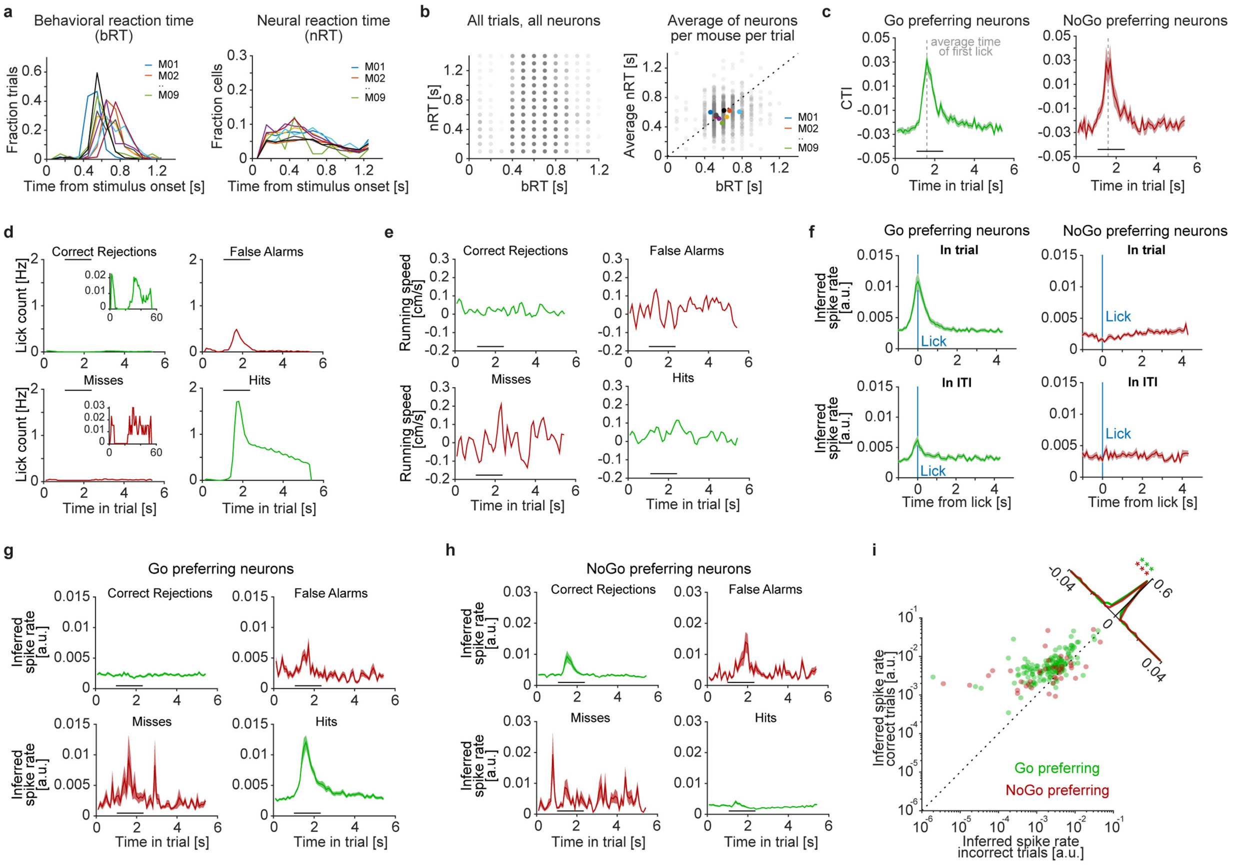
Extended Data Fig. 5 | Examples of single cells that became category-selective over the course of learning. **a**, Example Go category-selective cell (from the mouse shown in Fig. 2). Top left, position of the cell in the two-photon image. Scale bar, 30 µm. Bottom left, HLS map of the example region (hue: category identity of the presented stimuli; cyan, category 1; pink, category 2; lightness: response amplitude; saturation: selectivity for the stimulus category). Top middle, before learning (T1), Δ*F*/*F* traces aligned to stimulus onset. Scale bar, 100% Δ*F*/*F*. Grey bar denotes stimulus presentation, 1.3 s. Bottom middle, mean inferred spike rate per stimulus. Right, after learning (T5). **b**, As in **a** for an example NoGo category-selective cell. **c**, Mean inferred spike rate per stimulus for six further category-selective cells from different mice. **d**, Response amplitude during the first presentation of each stimulus, averaged across all category-selective cells at T5. Left, green: Go-preferring neurons; right, red: NoGo-preferring neurons. **e**, Top row, 40 by 40 pixel cropped images showing a Go category-selective cell in the averaged two-photon imaged field of views (pseudo-coloured). Bottom row, mean inferred spike rate of the response of the example cell to the presented stimuli at each imaging time point. **f**, **g**, As in **e**, but for different, NoGo category-selective cells.



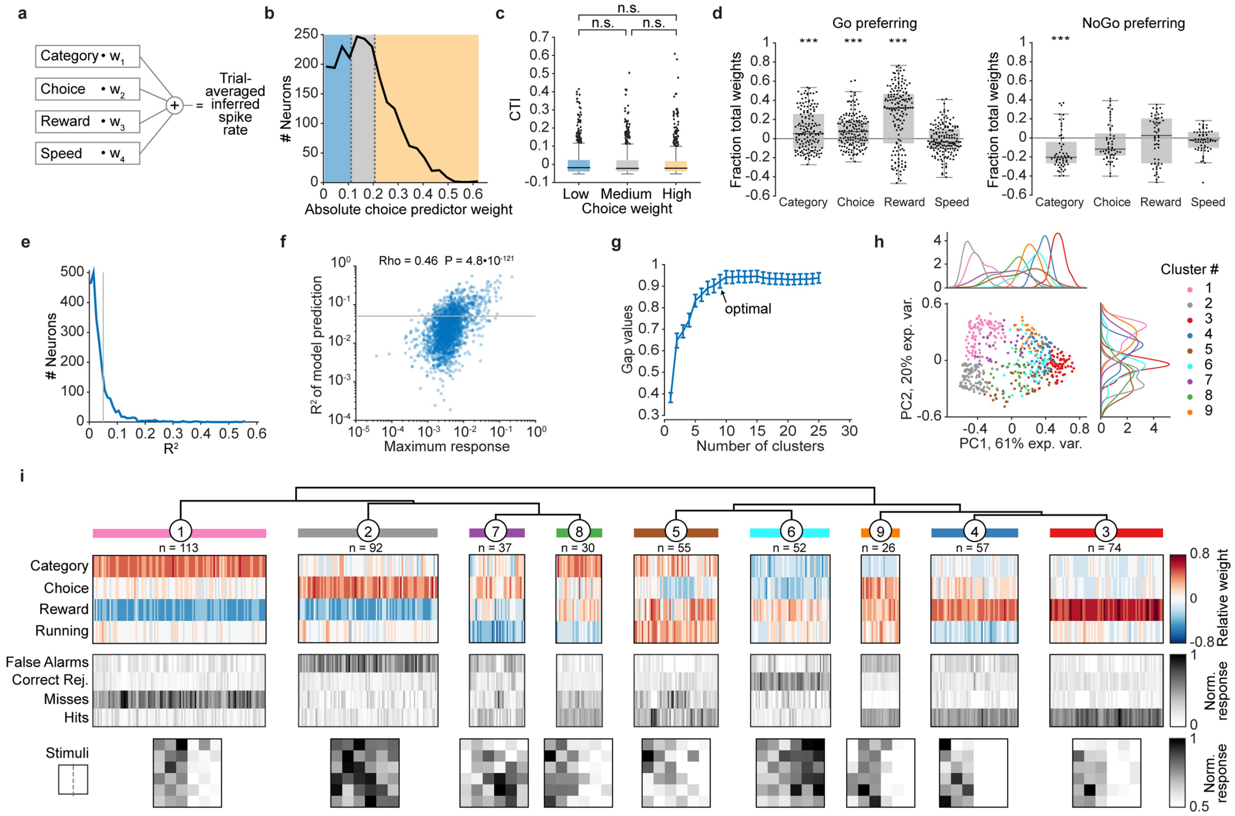
Extended Data Fig. 6 | Category-tuning index distributions of all recorded field of views at T1, T5 and T8 and the overlap of populations of category-selective cells. a, Category-tuning index before learning (T1, according to rule 1), after the mouse had learned rule 1 (T5, according to rule 1) and after it had learned to categorize stimuli according to rule 2 (T8, according to rule 2). Each imaging region is displayed individually. Individual cells are represented as dots. Only cells recorded at all imaging time points were included. Grey line indicates the threshold CTI value of 0.1 that was applied to classify cells as category-selective. **b**, Black, the fraction of overlap between category-selective groups found at T5 and T8 (Go stay/NoGo stay: Go/NoGo category-selective at both T5 and T8; Go switch: Go category-selective at T5 and NoGo category-selective at T8; NoGo switch: NoGo category-selective at T5 and Go category-selective at T8). Grey denotes 95% confidence intervals of chance population overlap (Methods). **c**, Go category-selective cells. Top row, Venn diagrams of the fraction of cells that were category-tuned only for rule 1 (R1), only for rule 2 (R2) or for both rules (area between R1 and R2). The highlighted part of the Venn diagram indicates which data are shown in the bottom row. Bottom row, mean stimulus response amplitude (inferred spike rate) after rule 1 (left) or rule 2 (right) was learned. **d**, As in **c**, for NoGo category-selective cells.



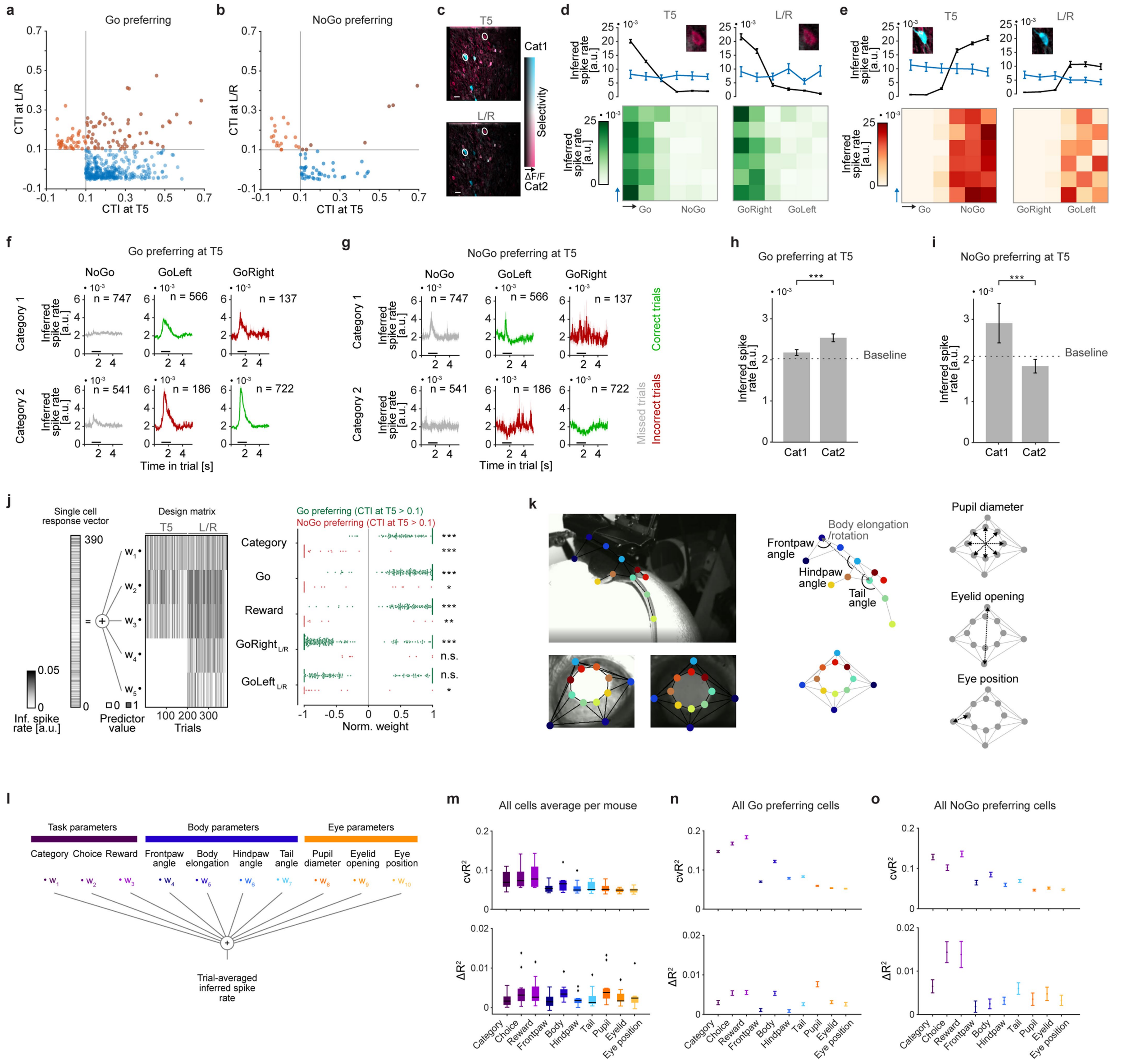
Extended Data Fig. 7 | Individual neurons follow characteristic time courses of acquiring selectivity. a, Left, scatter plot showing the difference in mean inferred spike rate between stimuli of the two categories, after learning the first rule (T5, *x* axis) and the rule-switch (T8, *y* axis) for individual Go category-selective cells at session T5 (blue) and T8 (orange). Right, histogram of the differences from unity of the distributions shown on the left. $P_{T5} = 1.5 \times 10^{-8}$, $P_{T8} = 3.5 \times 10^{-15}$, two-tailed WMPSR ($n_{T8} = 122$, $n_{T5} = 156$). **b**, As in **a**, but showing the relative spike rate difference (normalized by the sum of inferred spike rate to category 1 and 2 stimuli) for individual Go category-selective neurons at T5 and T8. $P_{T5} = 4.9 \times 10^{-21}$, $P_{T8} = 1.5 \times 10^{-18}$, two-tailed paired-samples *t*-test ($n_{T5} = 156$, $n_{T8} = 122$). **c**, As in **a**, but for NoGo category-selective cells at T5 and T8. $P_{T5} = 9.1 \times 10^{-5}$, $P_{T8} = 1.0 \times 10^{-12}$, two-tailed WMPSR ($n_{T5} = 57$, $n_{T8} = 70$). **d**, As in **b**, but for NoGo category-selective cells. $P_{T5} = 9.6 \times 10^{-8}$, $P_{T8} = 1.7 \times 10^{-21}$, two-tailed paired-samples *t*-test ($n_{T5} = 57$, $n_{T8} = 70$). **e**, Development of the spike rate difference up to T5, for individual Go category-selective neurons at T5. Before learning, baseline: T1. After learning the initial stimuli: T2. After learning categorization: T5. Grey lines denote individual neurons. Black line denotes the mean across cells. **f**, As in **e**, but for NoGo category-selective neurons. **g**, Schematic showing predicted time courses for the acquisition of reward/choice (RC) selectivity, and category selectivity according to each rule (R1, R2). These predictors were fit to the time courses of individual neurons using linear regression in **h**–**k**. **h**, Left, mean (± s.e.m.) predictor weight of T5 Go category-selective neurons. $P_{RC} = 3.2 \times 10^{-9}$, $P_{R1} = 2.0 \times 10^{-7}$, $P_{R2} = 0.12$, two-tailed WMPSR tests, Bonferroni-corrected for three comparisons ($n = 156$). Right, the predictor weights of individual neurons. Selectivity of Go-preferring neurons was best predicted by reward/choice, and also showed a category component. **i**, As in **h**, for T5 NoGo category-selective cells. $P_{RC} = 0.03$, $P_{R1} = 0.001$, $P_{R2} = 0.03$, two-tailed WMPSR tests, Bonferroni-corrected for three comparisons ($n = 57$). Selectivity of NoGo-preferring neurons corresponded best to the time course of acquiring category rule 1. **j**, As in **h**, for Go category-selective cells defined at T8 $P_{RC} = 0.03$, $P_{R1} = 0.003$, $P_{R2} = 7.8 \times 10^{-15}$, two-tailed WMPSR tests, Bonferroni-corrected for three comparisons ($n = 122$). **k**, As in **h**, for NoGo category-selective cells defined at T8 $P_{RC} = 0.09$, $P_{R1} = 0.52$, $P_{R2} = 8.0 \times 10^{-7}$, two-tailed WMPSR tests, Bonferroni-corrected for three comparisons ($n = 70$). The best predictor for both Go and NoGo preferring category-selective neurons after the rule-switch was the gradual acquisition of category rule 2.



Extended Data Fig. 8 | Relation between motor behaviour and neuronal responses of category-selective cells. a, Line histograms showing the count probability of behavioural (left) and neural (right) reaction times of individual mice. Behavioural reaction time (bRT) was measured as the time of the first lick after stimulus onset, neural reaction time (nRT) as the time of the neuronal response onset after stimulus onset. **b**, Left, scatter plot of bRT and nRT for every trial of every mouse in session T5. $P = 2.3 \times 10^{-13}$, rho = 0.08, Spearman's correlation ($n = 9{,}348$ measured reaction times). Right, grey circles: scatter plot showing the average nRT (that is, the nRT averaged across all Go category-selective neurons, but separated per mouse and trial) versus the bRT per mouse and trial. $P = 6.2 \times 10^{-6}$, Pearson's $r = 0.13$ ($n = 1{,}156$ trials). The density of grey circles is indicated by the colour intensity (alpha value). Coloured circles: the overall mean nRT and bRT of each mouse. $P = 0.51$, Pearson's $r = 0.26$ ($n = 9$ mice). Dashed line denotes the unity line. **c**, CTI of Go (left) and NoGo (right) category-selective neurons, calculated for every imaging frame individually. Data show the period from 1 s before stimulus onset to 3 s after stimulus offset. Grey dashed line denotes the average time of first lick. Black line denotes the average period of stimulus presentation. **d**, Mean lick frequency in session T5, grouped by trial outcome (hits, misses, correct rejections and false alarms). Insets show the same data with inflated *y* axis. Black line denotes the average period of stimulus presentation. **e**, As in **d**, but showing the average running speed. **f**, Inferred spike rate of Go (left) and NoGo (right) category-selective neurons aligned to the onset of lick-bouts. Top row, lick-bouts detected within a trial. Bottom row, lick-bouts detected in the inter-trial-interval. Data are mean ± s.e.m. **g**, Inferred spike rate of Go category-selective neurons in session T5, grouped by trial outcome (hits, misses, correct rejections and false alarms). Black line denotes stimulus presentation. Data are mean ± s.e.m. **h**, As in **g**, for NoGo category-selective neurons. **i**, Scatter plot showing the mean inferred spike rate in correct trials versus incorrect trials, for individual Go (green) and NoGo (red) category-selective neurons. $P_{\mathrm{Go}} = 1.0 \times 10^{-26}$, Pearson's $r_{\mathrm{Go}} = 0.72$, $P_{\mathrm{NoGo}} = 4.1 \times 10^{-5}$, Pearson's $r_{\mathrm{NoGo}} = 0.52$ ($n_{\mathrm{No}} = 156$, $n_{\mathrm{NoGo}} = 57$). Black line denotes the unity line. Line histogram shows the distribution of difference from unity separately for Go and NoGo-preferring neurons. $P_{\mathrm{Go}} = 6.6 \times 10^{-22}$, $P_{\mathrm{NoGo}} = 1.9 \times 10^{-5}$, two-tailed WMPSR ($n_{\mathrm{Go}} = 156$, $n_{\mathrm{NoGo}} = 57$).



Extended Data Fig. 9 | mPFC contains neural correlates of multiple task components. **a**, Linear regression model, fitting the trial-averaged inferred spike rates of individual neurons at T5 ('w_i' denotes the predictor weight; category predictor 0: category 1; 1: category 2; Methods). **b**, Distribution of absolute choice predictor weight of all observed neurons, divided into low, middle and high weight groups with equal numbers of cells. **c**, Box plots of CTI distributions for the choice weight groups in **b**. Boxes show the first to third quartile of the distributions, and black line denotes the median. There was no significant difference between the distributions, showing that category selectivity is not observed exclusively in highly choice-correlated cells. $P = 0.92$, Kruskal–Wallis test comparing all groups, chi-squared = 0.158, d.f. = 2. **d**, Relative weights of linear regression predictors (category identity, choice, reward and running speed) of Go and NoGo category-selective cells at T5. Left, category, choice and reward predictors show a significant deviation from 0. Right, only the category predictor shows a significant difference from 0. $P_{\text{Go-w1}} = 6.8 \times 10^{-5}$, $P_{\text{Go-w2}} = 2.3 \times 10^{-7}$, $P_{\text{Go-w3}} = 2.0 \times 10^{-14}$, $P_{\text{Go-w4}} = 0.17$, $P_{\text{NoGo-w1}} = 3.9 \times 10^{-5}$, $P_{\text{NoGo-w2}} = 0.03$, $P_{\text{NoGo-w3}} = 0.68$, $P_{\text{NoGo-w4}} = 0.11$, two-tailed WMPSR tests, Bonferroni corrected for four comparisons ($n_{\text{Go}} = 156$, $n_{\text{NoGo}} = 57$ cells). Grey boxes span the first to third quartile, black lines show the median. **e**, Distribution of R^2 values, black line at 0.05 denotes the cut-off for cells included in hierarchical clustering (resulting in 536 out of 2,306 neurons, largely excluding unresponsive neurons). **f**, Correlation of the R^2 value of individual cells and their maximum average response to correct or incorrect trials of either category. $P = 4.8 \times 10^{-121}$, rho = 0.46, Spearman's correlation ($n = 2{,}306$ cells). Grey line denotes the R^2 cut-off shown in **e**, which eliminated mostly unresponsive neurons. **g**, Gap statistic of hierarchical clustering for varying cluster numbers. Arrow denotes the optimal number of clusters (nine clusters; Methods). Error bars denote the standard error of the gap statistic value. **h**, Principal component analysis of model weights shows cluster separation along the major axes of variance. Line histograms show distributions per cluster along PC1 and PC2 separately. Individual neurons (dots) are colour-coded by cluster identity. **i**, Top, dendrogram showing cluster linkage. Second row, for each neuron, relative weights of model predictors in each of the nine clusters. Third row, for each neuron, normalized responses in the four different trial outcomes. Fourth row, per cluster, mean normalized response to every stimulus.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Category selectivity throughout the task change and contributions of task-relevant parameters and uninstructed movements to explained response variance. a, Scatter plot showing the CTI of Go-preferring neurons having a CTI > 0.1 in session T5 (blue) or session L/R (orange). Grey lines denote the CTI threshold used to determine category selectivity. **b**, As in **a**, for NoGo preferring cells. **c**, HLS maps of the example imaging region before (T5) and after (L/R) the task change (also shown in Fig. 4). Scale bar, 30 µm. White circles indicate example cells in **d** and **e**. Hue: preferred category; lightness: response amplitude; saturation: category selectivity. **d**, Example Go-preferring neuron. Top, inferred spike rate for stimuli ordered along the relevant dimension (black), or the irrelevant dimension (blue). Inset, section from the HLS map in **c** showing the example cell. Bottom, average inferred spike rate per stimulus. Data are mean ± s.e.m. **e**, As in **d**, but for a NoGo preferring example cell. **f**, Inferred spike rate of Go category-selective neurons (selected at T5), separated by stimulus/trial outcome combination in the left/right choice task. Top row, category 1 (GoLeft is correct). Bottom row, category 2 (GoRight is correct). Grey, missed trials, no reward. Green, rewarded trials. Red, unrewarded trials. Black line, stimulus presentation. In each panel, '*n*' indicates the total number of included trials (from nine mice). Data are mean ± s.e.m. **g**, As in **f**, for NoGo category-selective neurons (determined at T5). **h**, Category-selective neuronal responses, in absence of behavioural responses (missed trials in the left/right choice task). Inferred spike rate for each category presented in the left/right choice task, of Go category-selective neurons selected at T5. $P = 3.3\times10^{-8}$, two-tailed WMPSR ($n = 407$). Data are mean ± s.e.m. **i**, As in **h**, but for NoGo category-selective neurons (determined at T5). $P = 0.002$, two-tailed WMPSR ($n = 48$). **j**, Left and middle, schematic of the linear regression model, fitted to all trials of sessions T5 and 2AC combined. The average trial spike rate of each neuron was predicted by a weighted sum of the predictors: Category, Go, Reward, GoRight and GoLeft. Response vector and design matrix of example session. Right, significant normalized weights of all category-selective cells. $P_{\text{Go-w1}} = 4.6\times10^{-19}$, $P_{\text{Go-w2}} = 6.1\times10^{-37}$, $P_{\text{Go-w3}} = 5.4\times10^{-43}$, $P_{\text{Go-w4}} = 1.1\times10^{-23}$, $P_{\text{Go-w5}} = 0.58$, $P_{\text{NoGo-w1}} = 2.2\times10^{-6}$, $P_{\text{NoGo-w2}} = 0.003$, $P_{\text{NoGo-w3}} = 0.002$, $P_{\text{NoGo-w4}} = 0.40$, $P_{\text{NoGo-w5}} = 0.006$, two-tailed WMPSR tests ($n_{\text{Go}} = 407$, $n_{\text{NoGo}} = 48$). **k**, Left, example cropped image of the body-imaging camera and the eye-imaging cameras, with overlaid marker positions (tracked using DeepLabCut[41,42]). Middle and right, schematics defining body and eye parameters derived from the tracked markers. **l**, Schematic showing predictors and the linear regression model used to fit the cells' mean inferred spike rates per trial. **m**, Top, box plots showing the maximum predictive power (cvR^2) of each model predictor ($n = 9$ mice). Boxes show the first to third quartile; black line denotes the median. Bottom, box plots showing the unique contribution (ΔR^2) of each model predictor. **n**, Maximum predictive power (cvR^2, top) and the unique contribution (ΔR^2, bottom) across Go category-selective neurons ($n = 407$ neurons). Data are mean ± s.e.m. **o**, As in **n**, but for NoGo category-selective neurons ($n = 48$ neurons).

nature research

Corresponding author(s): Pieter Goltstein

Last updated by author(s): Jan 26, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavioral and imaging data were collected using custom written Matlab code that is available from the corresponding author on request. The custom behavioral training protocol was created using the publicly available psychophysics toolbox for Matlab as cited in the Online methods.
Data analysis	We used custom Matlab code to analyze the behavioral data and to process and analyze imaging data. For image processing and video tracking we also used publicly available video analysis, image registration and spike inference algorithms as cited in the Online methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the results of this study are available on: https://gin.g-node.org/sreinert/Category-learning_mPFC.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not use a power analysis to determine sample size before the experiment. The datasets were collected before analysis began and no data were added subsequently.
Data exclusions	One animal was excluded from all analyses of time point T8 and all analyses relying on chronically identified neurons, because optical access to the imaging region was lost in that animal at the last imaging time point (T8).
Replication	Aside from the replication of our main results in individual animals (n = 11, n = 9) as reported in the manuscript, no further measures were taken to ensure replication.
Randomization	Animals were randomly assigned to the categorization rule 'spatial frequency' or 'orientation'. There were no further experimental groups.
Blinding	We did not perform the analyses blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
□	☒ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female C57BL/6 strain Mus musculus (P63 - 82 at day of surgery) were obtained from the in-house breeding facility.
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	All procedures were performed in accordance with the institutional guidelines of the Max Planck Society and the local government.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

4 | Discussion

Forming and recalling categories is fundamental to our lives as they help us make quick decisions in unfamiliar environments. We share this ability with a wide range of vertebrate species, e.g. pigeons (Herrnstein and Loveland, 1964), rats (Vermaercke et al., 2014) and nonhuman primates (Freedman et al., 2001; Smith et al., 2010), and also invertebrate species, like crickets (Wyttenbach et al., 1996), jumping spiders (Dolev and Nelson, 2014) and honey bees (Benard et al., 2006). Depending on the structure of the categories and the underlying task, different networks of brain areas will engage to varying degrees in order to support the formation of a category (for review see Ashby and Ell, 2001).

One area of specific interest to category learning research is prefrontal cortex, as it is critical for explicit, rule-based categorization in humans (Milner, 1963) and a neuronal category representation, in the form of category-selective neurons, has been found in primates (Freedman et al., 2001). Prefrontal cortex is generally known to be involved in goal-directed behavior (Jacobsen, 1928), attentional processes (Fuster and Alexander, 1971) and flexible decision making (Winocur and Eskes, 1998). On the other hand it is also thought to play a role in long-term declarative memory (Simons and Spiers, 2003).

The goal of this thesis is to contribute to the understanding of category learning in prefrontal cortex from a systems neuroscience perspective. I investigated category learning in mice and followed learning-related changes to neuronal activity in medial prefrontal cortex (mPFC). First, I established that mice can learn rule-based categorization, generalize to novel stimuli and learn to regroup stimuli after a switch in rules. Hereby, I found that individual neurons in mPFC showed category-selective responses after learning. Further, this category selectivity was largely rule-specific and emerged gradually with learning. Lastly, I showed that the category representation in mPFC generalized across different motor requirements and reward contingencies.

4.1 Did mice learn rules for categorization?

Can mice learn a rule for categorization? The answer to this question needs to be 'yes' if we want to use the mouse model to understand neuronal circuits for explicit, rule-based category learning (see 1.3.2.2). Traditionally, explicit categorization was defined as forming a verbalizable rule that best distinguishes the categories (Ashby et al., 1998). This link to language defined it as an exclusively human ability. However, several key findings from human category learning studies were soon reproduced in non-human primates. Both humans and non-human primates showed a performance advantage in explicit over implicit categorization (Smith et al., 2010). Likewise, modeling of the category boundary supported selective attention to relevant features in explicit categorization (Smith et al., 2010). An independent study showed that non-human primates are able to flexibly switch rules for categorization (Roy et al., 2010). Together, those findings supported the hypothesis that monkeys can learn explicit categorization even though they lack language. This led to a re-consideration of the definitions of explicit and implicit category learning (Smith et al., 2010).

This updated definition, centered on selective attention to category-defining features, allowed for investigating explicit category learning in other species, like pigeons and rats. In contrast to humans and non-human primates, neither pigeons nor rats showed a difference in performance between explicit and implicit tasks (Smith et al., 2011; Vermaercke et al., 2014). This result has been used to argue that those species lack the explicit category learning system, and therefore solve every category learning problem with just one, implicit category learning strategy. However, more detailed investigations of categorization strategy in rule-based and information-integration categorization tasks in rats revealed that the animals did differentially weigh stimulus features in the rule-based task, but not in the information-integration task. This indicated that rats show a level of selective attention towards relevant stimulus features that is characteristic of explicit categorization (Broschard et al., 2019b).

While the study in chapter 3 lacked the direct comparison between rule-based and information-integration categorization in mice, I will discuss three measures - generalization, selective attention to the relevant stimulus feature, and rule-switching - that can give an indication of whether mice are able to learn a rule for categorization.

4.1.1 Generalization of novel stimuli

Generalization refers to the transfer of existing knowledge to a new problem (Estes, 1994). In category research, this transfer involves applying a learned classification to novel stimuli through retrieval and mapping (Shepard and Chang, 1963; Casale et al., 2012). Early on, Thorndike (1913) hypothesized that stimulus generalization is based on perceptual similarity (Gentner et al., 1993). Specifically, an exponential decay of generalization performance with increasing perceptual distance was proposed (Shepard, 1958; Gluck, 1991), in line with predictions from the exemplar theory of category learning (Medin and Schaffer, 1978).

This theory of generalization based on similarity suggests that there should not be an inherent difference between explicit or implicit category learning, as long as the underlying category structures and the similarity between trained and tested stimuli is identical (Thorndike, 1913; Gentner et al., 1993). However, human subjects showed a marked difference between rule-based and information-integration categories even if they had identical underlying structures. When the subjects were tested for generalization to perceptually very dissimilar stimuli, transfer was nearly perfect in the rule-based condition, but no evidence for generalization could be found in the information-integration condition, when carefully controlling for rapid learning (Maddox et al., 2005; Casale et al., 2012). This finding was in contrast to the similarity-based transfer hypothesis, but predicted by the COVIS theory of category learning (see 1.2.4.2; Ashby et al., 1998). According to COVIS, a rule-based category structure is learned by an explicit, hypothesis-testing system that forms a context-independent rule. This would allow for good generalization despite perceptual dissimilarity. Information-integration categories, on the other hand, are learned through an implicit, procedural learning system that maps a confined region of perceptual space onto a category identity. This region encloses the experienced stimuli and perceptually similar stimuli, but as soon as a novel stimulus is outside of the mapped space, generalization of the learned category distinction will not work. From this follows, that while both explicit and implicit category learning systems allow stimulus generalization based on the perceptual similarity (Maddox et al., 2005; Smith et al., 2015), only the explicit category learning system offers the similarity-independent ability to extrapolate a learned rule to an unfamiliar stimulus space (Casale et al., 2012; Smith et al., 2015).

After mice were trained on 18 stimuli (‘experienced stimuli’), I tested for generalization by presenting 18 novel stimuli. These were, similar to the study by Casale et al. (Casale et al., 2012), in an unfamiliar region of perceptual space with respect to the irrelevant stimulus dimension (see Chapter 3, Fig. 1i for stimulus dimensions). Further, since the novel stimuli varied along both stimulus dimensions, their similarity to the experienced stimulus set varied. In this generalization test, all mice performed equally well on the novel stimuli as on the experienced stimuli (Chapter 3, Fig. 1e,h). In addition, no relationship between stimulus similarity to the experienced stimuli and performance could be observed (see Chapter 3, Extended Data Fig. 2). Although, in mice, one trial learning of a stimulus association is rarely achieved with operant conditioning of visual stimuli (unlike classical conditioning of odorants or tastes; Welzl et al., 2001; Armstrong et al., 2006), a rapid incorporation of the novel stimuli into the learned classification could explain an average high performance on novel stimuli. In order to rule out the possibility that mice used the feedback they got to rapidly form an association with the novel stimuli, I separately quantified the animals' classification performance during the very first presentation of each stimulus. Importantly, even without any chance to learn through feedback, the mice showed no difference in their performance on experienced stimuli and on novel stimuli (Chapter 3, Extended Data Fig. 3).

In summary, mice showed the ability to generalize the learned rule beyond their experienced perceptual space. Further, no gradient of performance across the novel stimulus set could be detected. This indicated, that such generalization was not dependent on the overall similarity of the novel stimuli to the experienced set. Both results point to a level of generalization that is only enabled by rule-based category learning.

4.1.2 Selective attention based on the active rule

The major advantages of rule-based category learning compared to implicit category learning are the speed and flexibility of reacting to novel stimuli (see 4.1.1) and novel environments or contexts. In a new environment, the rules that apply to grouping and reacting to stimuli are often different. Upon a sudden change in environment, behavioral flexibility allows rapid remapping of reactions to stimuli that is in stark contrast to the time that would be required to individually re-learn each stimulus-response pairing (Roy et al., 2010). In the context of category learning, a hallmark of the flexibility achieved through explicit, rule-based learning is selective attention to the feature(s) relevant to the category distinction (Ashby et al., 1998).

As selective attention has long been characterized as a higher cognitive function of humans and non-human primates, it was debated whether rodents were also capable of perceptually separating and selectively attending to individual stimulus features (Seamans et al., 2008). Only recently, Broschard et al. (2019b) compared rats in a rule-based and an information-integration category learning task and modeled the influence of each stimulus feature on the decision boundary of individual animals. Hereby, the study showed that when rats performed the rule-based category learning task they differentially weighted the two stimulus features, with a higher weight for the feature relevant to the category distinction. This result supports the hypothesis that rodents are capable of selective attention (see also Carli et al., 1983; Robbins, 2002).

Similarly, the data on mice in my rule-based category learning paradigm did not show any influence of the irrelevant stimulus feature (see Chapter 3, Fig. 1j). The mice based their decisions only on the stimulus feature relevant to the active rule, indicating that the animals, indeed, perceptually separated and selectively attended to the individual stimulus feature.

Behavioral flexibility also manifests in the ability to rapidly remap stimulus responses upon a change in rules (Ashby and Spiering, 2004; Smith et al., 2012). Such rule-switches have typically been tested in mice by switching the stimulus modality that is relevant to the behavioral choice in a task (but see Biró et al., 2019 for rule-switches in the visual modality only). Mice were able to flexibly switch between the modalities when determining their response in task (Rikhye et al., 2018; Spellman et al., 2021). However, in the context of rule-based category learning the categories span a two- or higher dimensional space and a rule-switch typically requires a change in selective attention to the newly relevant stimulus feature (Heaton and Pendleton, 1981; Mansouri, 2006; Roy et al., 2010). In mice, flexible categorization has so far only been tested along a single dimension by shifting the threshold value determining the category identity of stimuli along that dimension (Zhong et al., 2019; Runyan et al., 2017; Kudryavitskaya et al., 2021). While mice successfully remapped the stimuli into the flexibly changing categories, these paradigms could only be compared to the results from primates in a limited manner, as they did not test for selective attention to individual stimulus features.

The stimulus space of the visual categories that I presented enabled a rule-switch from one stimulus feature (e.g. spatial frequency) to the other (e.g. orientation). The mice were able to learn the switch in rules and reassigned the stimuli to novel categories based on the newly relevant feature (see Chapter 3, Fig. 1c,g). Hereby, they learned the new category identities much faster than learning stimuli during the initial learning phase (see Chapter 3, Fig. 1f).

This increased learning speed indicated that the mice used selective attention to the relevant stimulus feature and shifted that attention to a new feature upon a switch in rules. Rule-switching in mice was not nearly as fast as in human (Robinson et al., 1980) or primate studies (1 trial; Roy et al., 2010). However, primates are typically trained on repeated rule-switches until such fast remapping is achieved before neural recordings are obtained. The mice, on the other hand, were only subjected to one rule-switch and hence had no opportunity to improve their remapping. Therefore, our data support the hypothesis that mice learned to apply a rule to categorize the visual stimuli and point towards a level of behavioral flexibility in line with rule learning findings (Havenith et al., 2018; Biró et al., 2019).

In summary, mice only weigh the relevant stimulus feature in their decisions upon learning to categorize visual stimuli. This is likely achieved by selectively attending to this stimulus feature. The animals further show much faster learning of the second rule compared to the first rule and rapidly remap stimulus-responses upon a rule-switch. Together, the results point to the hypothesis that mice benefit from the characteristic advantage of explicit categorization: behavioral flexibility through reassigning selective attention whenever contexts and, hence, the rules for categorization change.

4.1.3 Generalization of a rule overcomes previous stimulus associations

In the previous sections, I have discussed how our data support the hypothesis that mice can generalize a learned rule to stimuli they have not seen before. Further, I have highlighted that flexible reassigning of selective attention enabled mice to rapidly re-learn categories upon a rule-switch. However, it is the junction of those two characteristics that we apply in our daily lives. When we learn a new rule to group something, that new rule often conflicts with our previous knowledge. Nevertheless, we are able to disregard what we had learned before and, by reassigning our attention, apply the new rule to group objects we have so far only encountered with previous rules. In short, we can generalize a new rule, even if it is in conflict with prior knowledge. In humans, this ability is commonly tested using the Wisconsin Card Sorting Test in order to assess cognitive function(Heaton and Pendleton, 1981).

In my rule-based category learning paradigm, mice had shown both generalization of a learned rule to novel stimuli and the ability to learn a rule-switch. I therefore tested whether mice could apply the new rule even if it conflicted with associations from learning the first rule. After the mice had learned the second rule for categorization on a subset of stimuli, I added stimuli that the mice had only experienced using the first rule in a generalization test (see Chapter 3, Extended Data Fig. 3e for schematic). In this generalization test, mice only based their category decisions on the newly relevant stimulus feature and showed no influence of the feature that was relevant for the first rule (Chapter 3, Fig. 1j). More specifically, even on stimuli where the first rule predicted a different outcome than the second rule (i.e. conflicting stimuli), mice showed different behavioral responses after learning the second rule compared to during training with the first rule. These results indicate that, after the rule-switch, mice determined the categories of stimuli based on the second rule. This was especially striking on stimuli they had only experienced with the first rule and therefore could not use training to remap those stimuli. In other words, the mice generalized the new rule despite conflicting knowledge (see Chapter 3, Extended Data Fig. 3d,f).

Even though the study lacked the direct comparison between a rule-based and an information-integration category learning task, I found that mice generalize a learned rule, show selective attention towards a relevant stimulus feature that can flexibly change and can disregard prior knowledge in favor of a newly acquired rule. These three findings are characteristic of explicit category learning as it can be observed in humans and non-human primates. Therefore, I conclude that mice can indeed learn rules for categorization.

4.2 How does the experiment shape the result?

The goal of establishing rule-based category learning in mice was to study the function of prefrontal cortex with respect to such behavior. Before discussing the results of that investigation, I will first consider potential constraints of the chosen approach on behavioral and neuronal observations. Investigating category learning in an experimental setting required choosing a motivational incentive (water), a stimulus modality (visual stimuli) and a task structure (predominantly a Go/NoGo task) for the training. In general, all these aspects affect the behavior that subjects display in experiments (Guo et al., 2014), and likely different parameters will be optimal for different animal species. Hence, designing a task requires finding a balance between optimizing settings for the chosen animal model and retaining comparability across species. Beyond an influence on behavioral results, experimenters' choices on these task aspects can also differentially affect the observed function of a brain area in that context (Jourjine, 2017; Musall et al., 2019; Eiselt et al., 2021; Santiago et al., 2021). Below, I will discuss how my rule-based category learning paradigm could have shaped the results on a behavioral and neuronal level.

4.2.1 Effects of the task design on behavioral results

The majority of category learning experiments use operant conditioning to achieve the desired learning in subjects. Operant conditioning (Skinner, 1935) pairs a behavioral operation with an unconditioned stimulus (US) that has either positive or negative valence (i.e. a reward or punishment, respectively). Through repeated association with a reward or punishment, the operant behavior will be reinforced or discouraged. Designing an operant conditioning paradigm involves choosing the US, a stimulus modality and an operant behavior. Each of those choices will affect the behavior that subjects display in the task and, hence, can impact the results of a study.

The following behavioral parameters can be influenced by the task design: the welfare of the subject, learning speed and maximum performance that is reached after learning. In the following paragraphs, I will detail my choices on US, stimulus modality and operant behavior for the rule-based category learning paradigm and discuss their potential effects on the aforementioned behavioral parameters.

In both chapter 2 and 3, I chose appetitive reinforcement for operant behavior, that is presenting an intrinsically rewarding stimulus as the US, in response to the correctly displayed behavior (Dickinson and Balleine, 1994). Positive reinforcers given in human studies are usually snacks or monetary rewards, real or virtual (Levy and Glimcher, 2011). Working with animals, researchers typically administer food, e.g. grapes or juice for primates (Remington et al., 2012; Watanabe and Funahashi, 2015) and seeds, soymilk or water for rodents (Toth and Gardiner, 2000; Guo et al., 2014). In such animal studies, the reinforcement through food or water alone is often not strong enough to motivate participation. In order to make food or water a stronger reinforcer, animals are typically put on a restriction regime, limiting their food or fluid uptake to either a fixed amount (calories or volume) every day or to a certain level of weight reduction (Toth and Gardiner, 2000; Tucci et al., 2006).

First, I aimed to characterize the effect of food or water as a positive reinforcer with respect to animal welfare and task performance. Hereby, I found that food and water as motivational incentives, and the corresponding restriction regimes, differentially influenced the welfare of the mice and their learning speed in the task, but not their maximum performance. On average, water restricted mice showed mildly higher discomfort scores, evaluating their dehydration level and fur condition, than food restricted mice, even though all animals were kept on 85% of their pre-restriction weight. This could mean that water restriction is impacting the animals physiology more than food restriction (Hamilton and Flaherty, 1973) or that the discomfort evaluation was more sensitive to detect discomfort from water restriction, like dehydration. Specifically, the scoring of the fur condition might be affected by reduced grooming as a strategy of mice to save water (Ritter and Epstein, 1974).

Water restricted mice reached the learning criterion in a visual discrimination task significantly faster than food restricted mice, even though they on average performed fewer trials per training session. While the faster learning of water restricted mice could not be explained by parameters assessing task motivation, like per day relative weight loss and anticipatory licks, each water drop amounted to a larger proportion of the daily water consumption than a soy milk reward amounted to with respect to daily caloric intake. Potentially, the water drop was perceived as a larger reward than the soy milk drop and might have created a larger incentive for the water restricted mice to learn the discrimination task. Plateau performance after learning was not different between food- and water restricted mice. Based on these results, I chose to use water rewards for operant conditioning in rule-based category learning, keeping in mind that the motivational state could be a confounding factor for the observed learning speed.

As the second parameter, the stimulus modality for category learning, I chose vision. The main reason for that was comparability of behavioral and neuronal results to human and non-human primate category research. In principle, stimuli of all sensory modalities can be categorized (olfactory (Howard et al., 2009), auditory (Wyttenbach et al., 1996; Ohl et al., 2001), somatosensory (Rossi-Pool et al., 2016) and even abstract categories independent of a modality (Sorscher et al., 2021)). However, the visual modality is most predominantly used in category learning research in humans and non-human primates (Freedman et al., 2001; Smith et al., 2010). Especially the comparison of visual rule-based and information-integration tasks (see Chapter 1, Figure 1.1) has provided important insight into explicit and implicit category learning systems and has been used successfully in several species (Smith et al., 2012).

The key factor in rule-based and information-integration categorization tasks is that the two features (here orientation and spatial frequency) need to be separately attended to or integrated, respectively. Humans are able to do both (Smith et al., 2010), but for pigeons, for example, it is less clear whether they are able to separate the two visual features (Smith et al., 2011). Since rats show similar behavioral results to pigeons (Vermaercke et al., 2014) and category learning in mice has only been tested using one-dimensional auditory stimuli (Zhong et al., 2019), it is possible that mice cannot separate these visual features.

Could this choice for visual stimuli have negatively impacted mice's performance in the category learning task? In other words, could the choice for visual stimuli lead to results indicating mice do not have an explicit category learning system, purely because mice are not able to separate orientation and spatial frequency? Likely, only the observed learning speed was affected by my choice: mice needed many training sessions to learn the task (see Chapter 3, Fig. 1c). Learning speed in my experiments was in line with other observations of head-restrained discrimination training in the visual (Andermann et al., 2010; Histed et al., 2012), somatosensory (Guo et al., 2014), or auditory (Sanders and Kepecs, 2012) modalities, but was slower than training in olfactory discrimination tasks (Komiyama et al., 2010; Abraham et al., 2012). However, the high generalization performance (see 4.1.1) and strong indications for selective attention to relevant stimulus features (see 4.1.2) argue against a fundamental difference in category learning due to the lack of feature separation, despite slower learning.

Lastly, in addition to the choice of reinforcer and the sensory modality of the stimuli, the choice of operant behavior for a given task will be reflected in behavioral results (Guo et al., 2014). In the studies presented in this thesis, I used licking on one or two lick spouts (Go/NoGo design, Chapter 3; two-alternative choice design, 2AC, Chapter 2; respectively) as the operant behavior. Mice were able to learn both tasks with stable performances of 66-95% correct trials (see Chapter 2, Fig. 4a,b; Chapter 3, Fig. 1c). However, mice in the 2AC condition (Chapter 2) learned the visual stimulus discrimination task fast than mice in the Go/NoGo condition (Chapter 3). The former required on average ten days of training while the latter took on average 23 days to reach criterion. In the Go/NoGo task, animals readily learned the 'Go' response, that is, to lick in response to a rewarded stimulus. Therefore, the observed difference in learning speed was likely due to learning the 'NoGo' response, i.e. to withhold from licking in response to the non-rewarded stimulus. Rodents in Go/NoGo tasks are inherently biased towards licking and, hence, need extensive training to learn the appropriate behavioral inhibition to overcome impulsive licking (Schwarz et al., 2010; Guo et al., 2014), presumably due to the repeated positive reinforcement of the 'Go' response.

Aside from the specific difference in learning speed between the Go/NoGo and the 2AC task, head-fixed operant conditioning is slower than operant conditioning in freely moving animals. There are two possible explanations. First, the type of operant behavior during head-fixation could be more difficult to learn. Since operant conditioning positively or negatively reinforces a displayed behavior of a subject, it builds on the natural behavioral repertoire of the animal. Specifically for positive reinforcement learning, the more frequently a behavior is naturally used, the more often it can be reinforced. Hence, a more naturalistic operant behavior will be easier and faster to train. Often, like in tasks that involve navigating a maze, the operant behavior is akin to natural foraging behavior and will give plenty occasion for reinforcement (Heisler et al., 2015).

Second, head fixation itself causes stress for the animals. A study found that blood cortisol levels of head fixed mice were substantially elevated compared to control animals and only significantly decreased after ten days of daily habituation (Juczewski et al., 2020). For both reasons, head-fixed paradigms require extensive habituation and training that precedes learning of a sensory detection or discrimination task (Guo et al., 2014; Juczewski et al., 2020) and learning speed can still be slow. Importantly though, after learning, plateau performance and reaction time tend to not be affected (Abraham et al., 2012).

In summary, the choice of the US, stimulus modality and operant behavior each had an effect on the behavioral results of the studies in this thesis. Hereby, learning speed was affected the most by the difficulty in acquiring the operant behavior. Crucially, the plateau performance after learning was not influenced by any of the discussed parameters. It is important to keep in mind that the task design can be a confounding factor when comparing the learning speed across species. However, slower learning in mice did not per se impair comparability of the category learning results to humans and non-human primates, especially since mice showed important characteristics of explicit category learning while successfully learning the task (see 4.1).

4.2.2 Confounding effects of the task design on neuronal results

Similar to the considerations that task design impacts behavioral results (see 4.2.1), neuronal observations can also be confounded by the experimenter's choice of motivational incentive, stimulus modality and operant behavior in a learning paradigm. Task design can influence which brain areas are involved, the magnitude of the role of a brain area and the kind of variables that are represented by a brain area. Below, I will discuss possible confounding effects of my paradigm for rule-based category learning on neuronal recordings in prefrontal cortex and potential impacts on the conclusions of the study.

The involvement of prefrontal cortex in category learning is likely unaffected by the choice of stimulus modality, motivational incentive or operant behavior. In several species, PFC has been shown to be involved in visual and auditory category learning (Freedman et al., 2001; Gifford III et al., 2005) and even the learning of modality-independent scene categories (Jung et al., 2018). Further, category learning studies in humans, non-human primates and rodents typically use very different motivational incentives (see 4.2.1) and operant behaviors, like key or touch screen presses or even verbalization. Specifically considering rule-based category learning, as used in this thesis, PFC is necessary for the task performance (Dias et al., 1996; Rossi et al., 2007; Broschard et al., 2021), regardless of the presented rewards or required motor behaviors.

On the other hand, the choice of task design has a profound effect on the kind of representations that can be detected and how those representations can confound the results of a study. Prominently, a wide variety of parameters are represented in prefrontal cortex during goal-directed behavior (Miller and Cohen, 2001; Merre et al., 2021). In primate and rodents, PFC neurons show selectivity for almost any task-relevant parameter in an abundance of tasks (Asaad et al., 1998; Mansouri, 2006; Pinto and Dan, 2015). Therefore, if one task aspect is associated with another, for example a stimulus with a motor behavior, it can be challenging to disentangle individual influences on neuronal activation. Beyond representations of single parameters, mixed selectivity, i.e. the responsiveness of neurons to linear or non-linear combinations of parameters, is a key feature of prefrontal cortex (Rigotti et al., 2013; see 4.4.1 for a detailed discussion). Therefore, which task parameters are relevant in a paradigm, and how these parameters relate to each other, can impact what can be concluded from neural data.

There are two approaches that can be employed to investigate the function of any brain area. The first approach is to study one behavioral task (like here a Go/NoGo visual categorization task) and to try to capture all neuronal representations of the task as holistically as possible. It does not restrict the investigation to one representation of interest (for example category selectivity). This approach can help to gain understanding of how a brain area can solve a specific task and what computations might underlie that process (Nieh et al., 2021). The second approach focuses on one neuronal representation (e.g. category selectivity) and tries to minimize the confounding effect of other task-relevant parameters (like choice behavior or reward). In primate studies, animals are typically trained in 'Delayed-match-to-category' (DMC) tasks and category selectivity is quantified during a delay period between sample and test stimulus presentation (Freedman et al., 2001). Such a task design can help to distinguish neuronal representations of category from motor or reward-related activity (Freedman and Assad, 2016). However, in pilot experiments in mice, I found that the training times that mice needed to perform the task were unfeasible for chronic two-photon imaging. An alternative possibility is to test for category selectivity across different tasks or contexts in order to understand whether it generalizes across tasks.

I applied both approaches in order to characterize prefrontal cortex function during category learning. First, because in the Go/NoGo task one stimulus category was associated with a 'Go' response, and also a reward, and the other category was not, category, choice and reward were highly correlated task-parameters. However, in trials where a mouse made an error, this coupling broke down. I made use of such partial decoupling by building a generalized linear model (GLM) that aimed to predict trial-by-trial activity of individual neurons from task-relevant parameters. Applying this GLM analysis to recordings of trained mice, I found neuronal selectivity to stimulus category, choice, reward and mixtures of those (see Chapter 3, Extended Data Fig. 9). This approach reproduced the finding that PFC holds representations of several behaviorally relevant parameters. In the future, this could be used to characterize the emergence of individual representations through learning and identify potential interaction that could give insight about computations within PFC.

However, the GLM analysis could not definitively disentangle the influence of stimulus category and choice behavior on neuronal activity. Following the second approach, I asked how category selectivity of individual neurons would generalize across different motor requirements and reward contingencies. After training mice in the Go/NoGo task, I changed the design to a 2AC ('lick left'/'lick right') task. Hereby, I found that a large proportion of neurons showed selectivity for task parameters in a context-specific way, i.e., in only one of the tasks. Most prominently, groups of neurons in PFC were significantly modulated by 'Go' choice in the Go/NoGo task, whereas in the Left/Right task neurons also represented 'lick left' and 'lick right' specifically. Importantly though, PFC also contained a fraction of neurons, whose category selectivity generalized across the different tasks without a confounding influence of behavioral choice or reward (see 4.3.1).

In conclusion, the task design has a profound effect on the observed neuronal representations in prefrontal cortex, in line with prominent proposed models of PFC function (Duncan, 2001; Miller and Cohen, 2001). On the one hand, this can be a confounding factor when interpreting neuronal recordings of PFC during category learning (that I will discuss in the following section 4.3) and therefore needs to be kept in mind. On the other hand, characterizing PFC activity with such detail can reveal underlying computations and general mechanisms of PFC function (see 4.4).

4.3 Prefrontal cortex in rule-based category learning

I established a rule-based category learning paradigm for mice that allowed me to study underlying neuronal mechanisms. In neuroimaging studies, neuropsychological data from humans and electrophysiological data from non-human primates it has been established that a large variety of brain areas are involved in different category learning tasks (Ashby and Maddox, 2005) and at different stages within the category learning paradigm (Smith and Minda, 1998). That is, there is no single 'categorization' area, but rather diverse networks of circuits depending on the task at hand (for review, see Ashby and O'Brien, 2005).

For explicit, rule-based category learning, prefrontal cortical areas have been shown to play a crucial role. They hold category representations that are more abstract than in other areas (Freedman et al., 2001; Brincat et al., 2018) and without prefrontal cortex the explicit category learning system, specifically, is impaired (Milner, 1963; Broschard et al., 2021). However, studies from humans and non-human primates leave open questions of when and how prefrontal neurons acquire category-selective responses, how these computations arise from the different types of inputs that prefrontal cortex gets and whether category-selective activity in prefrontal cortex is necessary to perform learned categorization, i.e. is part of the semantic memory of the categories.

In the study in chapter 3, I aimed to characterize category-selective neuronal activity in prefrontal cortex of mice. Using chronic two-photon calcium imaging, I followed the activity of individual prefrontal neurons through category learning to gain insight about when and how category selectivity arises during learning.

4.3.1 Category selectivity after learning

To find out whether the mouse model can contribute to the understanding of category learning in the brain, I needed to establish whether there are commonalities between the human/primate brain and the mouse brain with respect to solving categorization tasks. In primates, electrophysiological studies discovered that roughly a third of all recorded neurons in dorsolateral prefrontal cortex (dlPFC) showed category-selective responses, after the animals were trained on a category learning paradigm (Freedman et al., 2001; Roy et al., 2010). These early findings put prefrontal cortex at the center of studies characterizing category representations and modeling circuit interactions (Miller et al., 2002; Brincat et al., 2018; Villagrasa et al., 2018). So far, it had not been addressed whether also mouse PFC has a role in category learning. No direct functional correspondence could yet be drawn between specific areas of primate PFC and mouse PFC (Carlén, 2017), as both are rather characterized by their heterogeneity (Merre et al., 2021).

In this thesis, I tested whether mouse medial prefrontal cortex (mPFC) is involved in rule-based category learning. Specifically, I investigated whether neurons showed category-selective activity in mice trained on a rule for categorization, as found in dlPFC of non-human primates. Indeed, after mice learned rule-based categorization, I detected neurons that showed stimulus-evoked activity to visual stimuli of one category over the other. On average, 10% of all recorded neurons in mPFC displayed such category selectivity following the active rule (see Chapter 3, Fig. 2g,h).

Due to the Go/NoGo task design, this observation of category selectivity could be confounded with neuronal representations of operant behavior and reward ((see 4.2.2)). In non-human primates this confounder is avoided by employing a DMC task design (Freedman and Assad, 2016). I used a switch from a Go/Nogo task to a left/right choice task to, similarly, disentangle the influence of motor and reward components on prefrontal neuronal activity. Hereby, stimulus category, operant behavior and reward were decoupled through the change in task. A unique contribution of the stimulus category on PFC activity was revealed through GLM analyses that included category identity, operant behavior and reward of both tasks, and even uninstructed, task-irrelevant behaviors (Musall et al., 2019) to predict neuronal activity. This unique contribution reflected neuronal activation that could only be explained by the presented categories and not by any of the potential confounders. More specifically, the GLM enabled identifying individual neurons whose selectivity generalized across both tasks. Strikingly, a fraction of the recorded PFC neurons was uniquely category-responsive (4.3%; similar proportion to rule encoding in mouse PFC, Rikhye et al., 2018). These neurons indicated the category of a presented stimulus in both the Go/Nogo and the left/right choice task, irrespective of the task context.

In summary, my rule-based category learning paradigm enabled reproducing a crucial finding from category learning research in non-human primates: category selectivity in prefrontal cortex. After learning, mouse PFC neurons show category-selective activity to presented visual stimuli, independent of displayed operant behavior and reward. The results are in line with findings of other functional features that mouse mPFC and primate dlPFC have in common, like the encoding of rules (Wallis et al., 2001; Rikhye et al., 2018), and point to an evolutionary conserved function of PFC in goal-directed behavior.

4.3.2 Rule specificity of category representations

Individual neurons in prefrontal cortex form representations of learned categories. But since we can learn, memorize and recall a huge number of categories, the question arises how the category representations relate to each other. In general, there are two hypotheses how such neuronal representations can be implemented: either, neurons could be part of multiple category representations, i.e. ‘categorization neurons’, or independent groups of neurons could represent different categories. These hypotheses likely have different implications for the neuronal mechanisms underlying category encoding and recall.

According to the first hypothesis, category-selective neurons are a set of neurons that get recruited with learning, together with other representations of task-relevant parameters (see Duncan, 2001; see 4.4.3). These neurons would be closely linked to the behavioral output, and therefore likely signal any learned category as long as it is relevant to the current behavior of a subject (Cromer et al., 2010). If prefrontal cortex indeed had such ‘categorization neurons’, they would remap their response behavior every time a stimulus-to-category mapping changed and, hence, different category identities became important after a switch in rules. This hypothesis is similar to global remapping of place cells in hippocampus upon entering a new environment (Muller and Kubie, 1987).

The second hypothesis proposes that, with learning, largely independent groups of neurons in prefrontal cortex gain category selectivity. This would enable learning many categories without overwriting previous knowledge and therefore facilitate memory and recall of the different categories (Roy et al., 2010). According to this hypothesis, likely, category-selective neurons are more closely linked to the sensory, here visual, input conveying information about the relevant stimulus features. In this case, different groups of neurons would become active whenever the stimulus-to-category mapping changes upon a rule-switch.

The rule-switch in my category learning paradigm aimed to identify whether prefrontal cortex rather held categorization neurons or independent representations. I trained mice on two categorization problems according to two rules, consecutively, and recorded activity of the same neurons through the learning period of both category rules. For each rule, a similar fraction of neurons (8-10% of all observed neurons) showed selectivity for the relevant categories (see 4.3.1; Chapter 3, Fig. 2g,h; Extended Data Fig. 6a). In order to relate those groups of category-selective neurons to each other, I calculated how many neurons were overlapping, i.e. were part of both groups. This revealed a difference between neurons selective for the ‘Go’ category and the ‘NoGo’ category. ‘Go’ category-selective neurons overlapped more than expected by chance, whereas the overlap of ‘NoGo’ category-selective neurons was within that expected range (12.6%, 3.9%, respectively; Chapter 3, Extended Data Fig. 6b). This finding indicated that a subset of category-selective neurons remapped their category responses in order to remain selective to the relevant feature for categorization (Chapter 3, Extended Data Figs. 5e, 6c,d). This is in line with the first hypothesis that neurons in PFC can encode multiple learned categories, hence that PFC holds categorization neurons. A similar result was obtained from non-human primates trained on two independent category sets, where a large fraction of neurons in PFC was found to represent both categorical distinctions (Cromer et al., 2010).

However, there are two important considerations to put my findings into context. First, the fraction of truly remapping category-selective neurons was likely smaller than quantified above due to confounding selectivity to the operant behavior or reward rather than stimulus category (see 4.2.2). Here, the time course by which a neuron acquired selectivity gave an indication what parameter influenced the neuron, since the operant behavior was learned earlier than the categories (see Chapter 3, Extended Data Fig. 7g). PFC held both, neurons that reflected learning of choice behavior, but, importantly, also remapping categorization neurons that acquired selectivity following the time course of category learning of each rule (see Chapter 3, Extended Data Fig. 7h,j).

Second, even though some category-selective neurons remapped to the novel categories, the majority of neurons did not. The neurons that did not overlap between the category-selective populations showed no stimulus responsiveness in the respective other condition, indicating that those neurons exclusively represented one learned set of categories (Chapter 3, Extended Data Fig. 6c,d). This finding is in line with results from non-human primates trained to switch between two rules for categorization (Roy et al., 2010). In this study, independent groups of neurons represented the categories that followed orthogonal rules on the same stimulus set.

What do the different observations imply for category representations in PFC? Likely, the two hypotheses are not mutually exclusive, but rather reflect the ends of a spectrum. That is, prefrontal cortex shows both categorization neurons and independent representations, while their respective contribution depends on the context of the task. If a subject trains two independent categorization problems simultaneously, like in Cromer et al. (2010), the categories do not conflict each other and, hence, can be encoded by categorization neurons. On the other hand, if stimuli change their category identity, as with the rule-switch (Roy et al., 2010), independent groups of neurons resolve the conflict. In line with a more graded theory, in the rule-switch condition also some neurons were selective for both categorical distinctions (see Extended Data Fig. 6b; 7.1% in Roy et al., 2010), and Cromer et al. (2010) also observed neurons that were exclusively selective to one categorical distinction.

In summary, in prefrontal cortex one set of category-selective neurons flexibly remap their selectivity to encode the behavioral relevance of the categories and another, larger set of neurons show rule-specificity in their representation of learned categories. Taken together, neither hypothesis – categorization neurons or independent representations – can be excluded. Rather, in prefrontal cortex a mixture of both can be observed, with the balance between them depending on the structure of the tested categories. Further work is needed to better characterize the relation and stability of category representations in mouse prefrontal cortex. For example, chronically recording neurons while mice perform another rule-switch, i.e. going back to the first learned rule, could help to better understand the kind of encoding of such category-selective neurons, their potential role in category memory and implications for the underlying circuitry.

4.3.3 Just broad visual tuning?

Could the observed category selectivity reflect the activity of visually driven neurons, independent of any category learning? Neurons that are broadly tuned to the presented stimulus features, orientation and spatial frequency, could appear category-selective. In order to answer this question, it needs to be addressed whether prefrontal cortex, as a target of direct inputs from visual areas (Pandya and Yeterian, 1990), per se displays tuning to visual features of the presented stimuli. In primary sensory areas, tuning to sensory features like visual orientation (Hubel and Wiesel, 1962) or auditory frequency (Phillips and Irvine, 1981) and multi-dimensional tuning (Jimenez et al., 2018) is common. If such responsiveness can be found in PFC, neurons with broad visual tuning to a combination of orientation and spatial frequency might be falsely classified as category-selective, because their response pattern existed independently of the learned categories.

Unlike in sensory cortical areas, there are few studies of sensory feature tuning in the absence of goal-directed behavior in the prefrontal cortex of any species. There is some evidence from object recognition tasks in primates that prefrontal cortex neurons respond to novel visual objects, and that with increasing experience, and hence familiarity of the object, fewer neurons respond with narrower tuning curves (Rainer and Miller, 2000). However, when comparing activity to the same visual stimuli before and after learning a visual working memory task, Qi et al. (2011) have observed more neurons being recruited after learning, but on average a decreased stimulus selectivity. These contrasting findings could be due to different measurements of neuronal activity and selectivity, but also potentially due to a task-dependency of the recruitment of PFC neurons. In non-human primates learning stimulus discrimination (Cromer et al., 2011) or categories (Antzoulatos and Miller, 2011), category-selective neuronal activity in PFC arose with learning (and abstraction of the category). These finding argue against preexisting tuning to underlying features and rather for a learned representation in PFC.

In the mouse, visual tuning to orientation and spatial frequency, as in visual cortical areas, has not been characterized in prefrontal cortex. Two observations in my experiments indicate that category selectivity in mouse PFC was not due to preexisting visual tuning. First, I compared category selectivity before and after learning. For this analysis, I recorded activity from PFC neurons in naïve mice to quantify 'baseline' visual tuning properties. Hereby, I presented all stimuli that would later be part of the categorization task. In this baseline session, hardly any neuron showed category selectivity (Chapter 3, Fig. 2g,h; Extended Data Fig. 6a). Likewise, neurons I identified as category-selective after learning did not show selectivity for categories, individual stimuli or visual features during the baseline session (Chapter 3, Fig. 2d; Extended Data Fig. 5).

Second, as described in detail in section 4.3.2, a fraction of category-selective neurons remapped their selectivity upon the rule-switch in order to signal the newly relevant categories (Chapter 3, Extended Data Fig. 6b-d). Such remapping required a significant change in their tuning to underlying visual features across a brief learning period. If the observed category selectivity was due to preexisting visual tuning in PFC, this tuning would likely not be flexible enough to support such a rapid change.

In summary, chronic two-photon imaging of calcium activity in individual prefrontal neurons allowed for comparing response properties before and after learning. While neurons showed category selectivity after learning, during baseline sessions no such responsiveness could be observed. This finding suggests that the category representation was built up with learning and not due to broad, preexisting tuning to the visual features in the task.

4.3.4 Category representation as a part of memory

Is the category representation in mouse PFC part of long-term memory for the learned categories? While the role of prefrontal cortex in coordinating goal-directed behavior is undoubted, it is less clear whether PFC is also involved in the storage of long-term memories (see also 4.4.3). One key implication for PFC circuitry would be that, with learning, lasting representations are formed that encode relevant parameters and that can be recalled when necessary. A possible mechanism for such a formation could be changes in connectivity of PFC neurons to their inputs from other brain areas, for example sensory inputs, or from within PFC. An alternative hypothesis is that PFC is not part of long-term memory and does not form representations through connectivity changes, but rather flexibly encodes task variables through sustained firing (Duncan, 2001; Miller et al., 2002).

Following the time course of emerging category selectivity in prefrontal cortex can give an indication how fast the representation is acquired and how stable it is across learning. In non-human primates, category learning studies typically only record neurons in a single training session (Antzoulatos and Miller, 2011). Such a time frame is not sufficient to determine stability of the representation and hence to distinguish between the hypotheses above. In contrast, the paradigm of chronic two-photon calcium imaging in mouse mPFC that I employed allowed for investigating the dynamics of the acquisition and the stability of category selectivity.

I observed mice from the state of a naïve performer, that has never been trained on a categorization problem, to an expert performer. Hereby, identified category-selective neurons could be investigated in all prior training stages. The two groups of category-selective neurons – 'Go' and 'NoGo' category-selective – on average showed different dynamics of acquiring category selectivity (Chapter 3, Fig. 3c,d). The 'Go' category-selective neurons displayed selective responses already after stimulus discrimination learning, i.e. before proper category learning had started. Further, on average they remained selective through the entire behavioral training, including the rule-switch. In contrast, 'NoGo' category-selective neurons gradually increased in their selectivity with category learning. This result pointed to a difference between the two groups of neurons.

However, for the two following reasons it was necessary to investigate the time course of individual neurons rather than their average. First, as discussed in section 4.3.2, the average time course was confounded by neurons selective to operant behavior and reward rather than stimulus category. Second, an average time course that appears like a gradual increase of selectivity could be composed of individual neurons that gain and lose selectivity in an instantaneous fashion, rather than gradually. I found that in both 'Go' and 'NoGo' category-selective groups, there were neurons whose time course of selectivity acquisition was best explained by the gradual acquisition of selectivity for a learned rule (Chapter 3, Extended Data Fig. 7g-k).

This finding is in line with the gradual increase in category information characterized in primate prefrontal cortex during the learning within one training session (Antzoulatos and Miller, 2011). While in primates the categories were learned - and thus also category selectivity was acquired - within few trials, the increase in category selectivity in mouse prefrontal cortex was observed over several days of training for each of the learned rules. This adds to the primate finding that the acquired selectivity is not due to the recruitment of a new set of neurons within each training session (Duncan, 2001), but rather a more stable component in the neuronal representation.

In summary, neurons in mouse prefrontal cortex acquire category-selective responses gradually over several days during the learning period. This observation of gradual emergence of category selectivity lays the groundwork for further experiments to probe the stability of the representation - i.e. probing the neuronal responses over a period of stable categorization behavior - and future perturbation experiments to determine whether there is a causal implication of category representations in PFC in categorization behavior. Such experiments will help to further disentangle the role of PFC in the long-term memory of categories and rules from its role in orchestrating goal-directed behavior (see 4.4.3).

4.3.5 PFC as part of a category learning network

I have so far discussed the category learning behavior in mice and the investigation of category representations in the brain, focusing on prefrontal cortex. However, from prior category learning research - in humans, primates and rodents - it is clear that PFC is by far not the only brain area involved in category learning. Rather, varying networks of brain areas interact depending on the category learning problem at hand. Therefore it is important to consider the function of prefrontal cortex in the context of findings from other brain areas, possible connections and proposed interactions.

Two neurocomputational models that I have described earlier (see 1.2.4.3; Knoblich et al., 2002; Villagrasa et al., 2018) aim to devise a mechanism by which category selectivity arises in prefrontal cortex. Both predict, based on electrophysiological observations in inferotemporal cortex (ITC) and prefrontal cortex of primates, that neurons in PFC acquire category selectivity through Hebbian plasticity at the connections between ITC and PFC. Input from object- or shape-selective neurons in ITC is hereby combined with information about what is relevant to the current task, for example through feedback about reward (a 'teaching signal'). This combination of inputs onto prefrontal neurons are suspected to drive plasticity. While Knoblich et al. (2002) do not speculate on how the latter information arrives in prefrontal cortex, Villagrasa et al. (2018) propose a connection from striatum via thalamus to prefrontal cortex to relay a teaching signal.

The choice- and reward-related activity I observed in mouse PFC (see Chapter 3, Extended Data Figs. 8,9; 4.2.2) could be a form of task-modulation that is necessary to drive plasticity between visual processing areas and prefrontal cortex. Both choice- and reward responsiveness were present in individual neurons of PFC as soon as the mice acquired basic task knowledge (T2, stimulus discrimination task; see Chapter 3, Extended Data Fig. 7). Such responses could be a teaching signal that, through changes in connectivity within PFC, gets combined with sensory input and therefore leads to a gradual acquisition of category selectivity. According to the model by Villagrasa et al. (2018), striatum learns stimulus-response associations through synaptic changes between ITC and striatum. This learned information gets relayed via thalamus to prefrontal cortex. The model circuitry is based on connectivity in the primate, but the key connections have also been described in the mouse (Pan et al., 2010; Hintiryan et al., 2016; Kuramoto et al., 2016). Even though ITC, as defined in primates, is not as clearly distinguished in the mouse, the mouse visual system also follows the broad distinction of dorsal and ventral stream (Wang et al., 2012). Areas associated with the ventral stream, like postrhinal cortex (POR), show functional similarities to ITC with respect to object sensitivity (Furtak et al., 2012) and also do connect to mouse prefrontal areas (Hwang et al., 2018). Further, context modulation of PFC neurons through thalamus has been observed in mice that learned different task rules (Rikhye et al., 2018). Such modulation could be due to the proposed connection between striatum and prefrontal cortex and could also be involved in category learning tasks.

In summary, the gradual acquisition of category selectivity and the representation of other task-relevant parameters are in line with predictions from network models of prefrontal cortex in category learning. Even though these models were based on results from studies in non-human primates, several proposed connections are conserved in the mouse.

The mouse model therefore provides several approaches to test predictions from elaborate network models. First, the task design that I developed allows for the recording of activity from a variety of brain areas throughout category learning. With chronic two-photon calcium imaging, activity in ventral visual stream areas can be observed (Goltstein et al., 2021). Such an investigation could potentially reproduce and expand upon results from primate area ITC. Second, through a task design that temporally separates stimulus presentation and reward, the precise timing of PFC neuronal activity can help elucidate what parameters modulate PFC neurons at what point in the training. With such design, possibly a key feature of the striatum-PFC interaction model could be reproduced – the shift in response timing from reward-related activity to stimulus-related activity that happens in primate PFC with learning (Pasupathy and Miller, 2005; Antzoulatos and Miller, 2011; Villagrasa et al., 2018). Further, connections between brain areas like ventral visual stream area POR and striatum or PFC can be precisely traced (Pan et al., 2010; Oh et al., 2014), forming a clearer picture of the possible underlying network interactions. This connectivity information can also be used to target projection specific investigations of activity, like imaging of synaptic inputs or outputs from downstream or upstream areas, respectively. Such information can help determine how different kinds of task-relevant information are conveyed across and integrated within brain areas and hence fill in the gaps in the network models of category learning.

4.4 A general theory of prefrontal cortex function?

Can the findings about prefrontal cortex activity in category learning help understand how this brain region generally functions? In other words, are there motifs in the involvement of prefrontal cortex in category learning that can be extrapolated to understand prefrontal computations in general?

So far, I have discussed the implications of my results on category learning in the brain and the role of prefrontal cortex during categorization. In the last section of this discussion, I will consider aspects how the results from prefrontal cortex during category learning can be integrated with prior research in order to understand if there is a general theory of prefrontal cortex function.

4.4.1 Mixed selectivity

Already early on, researchers characterized the activity of prefrontal cortex neurons in different tasks and contexts, and described neuronal selectivity to various task-relevant parameters (see 1.2.3.2). Historically, studies mainly looked for selectivity in a 'classical' sense, that is, preferential responses to an individual parameter that can be isolated from other factors (Rigotti et al., 2013). Neurons selective to various parameters were identified, such as neurons showing selectivity for a stimulus during working memory (Fuster and Alexander, 1971), or rule- (Wallis et al., 2001; Rikhye et al., 2018) and category selectivity (Freedman et al., 2001), but also reward- (Niki and Watanabe, 1979; Lak et al., 2020) and choice-selective responses (Asaad et al., 1998; Pinto and Dan, 2015; Lui et al., 2021). These findings showed the abundance and diversity of neural encoding of task parameters in prefrontal cortex, but also highlighted the difficulty in identifying a unified principle of PFC function.

However, it became clear that characterizing neurons by their selectivity to one parameter did not capture the complexity of neuronal activation. The majority of neurons in prefrontal cortex is driven by linear or non-linear combinations of two or more task parameters. This type of encoding was termed 'mixed' selectivity (Rigotti et al., 2013). In several species, studies have reported linear or non-linear mixed selectivity in prefrontal cortical neurons. Examples that have been described are context- or rule modulation (Mansouri, 2006; Zheng et al., 2021), integration of target location and task epoch (Parthasarathy et al., 2017) or higher dimensional interactions of task parameters (Balaguer-Ballester et al., 2011; Kobak et al., 2016).

In chapter 3, I mainly characterize neurons that show selectivity to learned categories, i.e. classically selective neurons. By focusing on disentangling stimulus category, choice and reward contributions, I describe neurons in PFC that are uniquely modulated by the presented category (see Chapter 3, Fig. 4h). The linear regression analysis that I used also revealed neurons that showed classical selectivity for other parameters, like reward or motor activity. Importantly, the majority of the neurons whose activity the regression model could capture (Chapter 3, Extended Data Fig. 9e), were best predicted by a linear combination of two or more task-relevant parameters. Most prominently, many category-selective neurons were also influenced by behavioral choice (Chapter 3, Extended Data Fig. 9d) and a large number of neurons were driven by combinations of category, choice and reward (Chapter 3, Extended Data Fig. 9i), hence showed linear mixed selectivity. Both, the observed classical and mixed selectivity, are in line with encoding of task-relevant parameters that has been previously characterized (Pinto and Dan, 2015; Lak et al., 2020).

There were, however, also neurons that could not be predicted well with the linear regression model (Chapter 3, Extended Data Fig. 9e). This could be due to several reasons. A straightforward explanation for a low prediction performance could be that those neurons were unresponsive during the task or their responses did not exceed the level of noise. Indeed, model performance correlated with a neurons detected response amplitude (Chapter 3, Extended Data Fig. 9f). However, another reason for a low prediction for some neurons could be that the model did not contain information that would be necessary to explain their activity. First, that could be precise temporal information about the task parameters and neuronal activity. Using event kernels to predict the neuronal activity across each trial rather than trial averaged activity (Runyan et al., 2017) could add necessary detail to a model to explain activity of prefrontal cortex neurons better. Second, a parameter of modulation by the activity of other PFC neurons could be included to capture neuronal coupling within prefrontal networks (Runyan et al., 2017).

Lastly, an important consideration is that the linear regression analysis was only able to characterize linear mixed selectivity. Non-linear mixed selectivity could not be fitted, because I did not include non-linear combinations of task parameters into the model. Primate prefrontal cortical neurons have been shown to be modulated by non-linear combinations of task parameters (Rigotti et al., 2013; Parthasarathy et al., 2017). In the mouse, non-linear encoding of learned task-relevant parameters has been described in the hippocampus (Nieh et al., 2021), but whether prefrontal cortex in this species employs such coding is yet unknown. Therefore, there is a possibility that a fraction of neurons was not fit well by my GLM due to non-linear mixed selectivity.

Mixed selectivity could be a general computational mechanism that the mammalian prefrontal cortex employs to efficiently solve complex tasks and to remain flexible enough to react to rapid changes in behavioral requirements. Representing combinations of parameters at the level of individual neurons is efficient because it increases the information that can be encoded in neuronal populations compared to classical selectivity (Fusi et al., 2016). At the same time, a simple linear read-out downstream is sufficient to retrieve information about any of the task-relevant parameters (Rigotti et al., 2013; Fusi et al., 2016). So far, it remains an open question how such mixed selectivity is computed by PFC neurons and what that implies for the stability of the acquired representations. It is possible that inputs to PFC from different brain areas convey different task-related information, for example that sensory areas like ITC convey information about the presented stimuli, striatal connections via the thalamus convey the task context and inputs from the ventral tegmental area (VTA) information about reward (Schultz and Dickinson, 2000; Han et al., 2017). Mixed selectivity could then emerge from PFC neurons flexibly connecting to inputs from multiple areas or from connections to classically selective neurons within PFC (Pinto and Dan, 2015).

By now, evidence has accumulated that such mixed selectivity is not unique to prefrontal cortex but also exist in other brain regions. Mixed selectivity has been described in posterior parietal cortex (Raposo et al., 2014), subiculum (Ledergerber et al., 2021) and even primary visual cortex (Keller et al., 2012; Saleem et al., 2013). This indicates that mixed selectivity is likely a general aspect of neuronal function across the cortex rather than specific to prefrontal cortex, further highlighting the importance of such a computational motif.

Taken together, the finding that prefrontal cortex forms classically and mixed-selective representations of task-relevant parameters allows for further investigations to elucidate prefrontal computations. Specifically, neuronal responses in this category learning paradigm can be characterized in more detail, by asking whether mouse PFC neurons also show non-linear encoding of task parameters, similar to primate dlPFC. Combining single neuron analyses, like described in this thesis, with analyzing population dynamics can reveal how mixed selectivity on the level of individual neurons gets integrated into population level representations. Importantly, in mice it is possible to manipulate specific inputs to prefrontal cortex to probe how PFC neurons gain mixed selectivity. This can help to understand how individual neurons receive and integrate various inputs about stimuli, choices and outcomes and hence contribute to higher dimensional representations and dynamics that are relevant for goal-directed behaviors.

4.4.2 Specificity to behaviorally relevant parameters

A striking aspect of prefrontal cortex activity is its specificity to encoding behaviorally relevant variables. Sensory cortical areas represent features, e.g. visual, auditory or somatosensory, of our environment. By now it is established that sensory areas show modulation of their activation based on attention (Corbetta et al., 1990; Maunsell and Cook, 2002) or the state of the animal

(Goltstein et al., 2015), and that learning a behavioral relevance can change those representations (Poort et al., 2015; Goltstein et al., 2018b). Nevertheless, representations of behaviorally irrelevant features (Poort et al., 2015) and selectivity independent of any training context (Hubel and Wiesel, 1962) can also be observed.

In contrast, prefrontal cortex seems to prominently represent what is relevant to an organism, regardless of any sensory modality (Rikhye et al., 2018; Zheng et al., 2021). Results from early studies describing persistent activity during working memory tasks already found single neurons selectively activated by a relevant task variable (Fuster and Alexander, 1971). Similarly, in both primate and rodent prefrontal cortex, neurons show selectivity to learned, relevant task rules (Wallis et al., 2001; Rikhye et al., 2018). Rule-based categorization is a paradigm that especially highlights this aspect of prefrontal cortex because the same input from the same modalities can be presented in different contexts, where only the task relevance of individual stimulus features changes. Neurons in primate PFC represent the categories that the monkey discriminates at any given time. Independent groups of neurons were found to be category-selective for each active rule (Freedman et al., 2001; Roy et al., 2010).

The results presented in chapter 3 support this finding (see 4.3.2). Since I recorded neuronal activity in mice before they learned any task-relevance, baseline prefrontal cortex representations of stimulus features could be assessed. In these baseline recording sessions, prefrontal neurons did not represent the stimulus features (Chapter 3, Fig. 2b,d). None of the presented visual features (grating orientation and spatial frequency) drove PFC activity of untrained mice before relevance was assigned to them. Further, the switch in rules for categorization revealed that hardly any PFC neuron showed category selectivity for the categories that followed the inactive rule, i.e. category-selective neurons ceased their activity as soon as those categories lost relevance or remapped to the newly relevant feature (Chapter 3, Fig. 3d,e).

The specificity to task relevance is in line with the adaptive coding model of prefrontal cortex (Miller and Cohen, 2001; Duncan, 2001; (see 1.2.3.2)). According to this theory, PFC neurons have the ability to encode whatever becomes relevant to an individual and to flexibly change their code upon changes in behavioral requirements. A key implication of the model is that each neuron gets diverse inputs about sensory, motor and reward features, but irrelevant inputs get weakened while task-relevant inputs increase in strength. Through such selective weakening and strengthening, tuning of PFC neurons for relevant features sharpens. With such encoding, PFC could orchestrate attentional modulation of other brain areas in order to drive goal-directed behavior, i.e. up- or downregulate activity in other brain areas depending on the task (Miller et al., 2002). This proposed aspect is often referred to as cognitive control. Theories of visual attention exemplify how PFC could enact such control over other brain areas: excitatory top-down inputs from PFC onto visual cortex neurons bias the activity of neurons tuned to relevant features and, through mutual inhibition, the activity of other neurons will be suppressed (Desimone and Duncan, 1995; Miller et al., 2002). Miller et al. (2002) further suggest that representations of task-relevant parameters in PFC connect with each other, forming a ‘model’ of the current task. PFC holding such a model would mean that partial information, such as a sensory cue, can activate the entire PFC network for the task (Hebb, 1949), enabling it to, in turn, activate other necessary components in different brain areas to elicit an appropriate reaction. In category learning theories (especially Ashby et al., 1998), this specificity to task relevance in prefrontal cortex is suggested to be a key component of active hypothesis testing processes that serve to discover relevant rules in the explicit learning system.

So far, it is unclear how task specificity is computed in prefrontal circuitry during learning and hence what underlies hypothesis testing on a cellular basis. One possibility is that the context information that reaches PFC through inputs from thalamus (Rikhye et al., 2018) modulates prefrontal representations based on their relevance. Prefrontal cortex could then internally connect these representations and form a model of the current task. Another open question is, whether irrelevant feature information - although not directly encoded in individual prefrontal neuron representations - remains present in the population, and for example could still be decoded. This could be a requirement in order to maintain the flexibility to rapidly react to changes in task or context. Further work on investigating the activation of prefrontal cortex and its inputs during goal-directed behavior and learning will be necessary to elucidate how prefrontal cortex implements one of its central features.

4.4.3 A dual role of PFC?

The final aspect that I want to discuss is, whether there is a role for PFC in long-term memory (see 1.2.3.2). It is undebated that PFC is involved in organizing goal-directed behavior. Mixed selectivity and the specificity to behavioral relevance, as described above, could be the core mechanisms for such PFC function and form a complete picture (Duncan, 2001; Miller et al., 2002). Alternatively, PFC additionally plays a role in long-term memory, as was hypothesized by Fuster (2003). One key assumption for a role of PFC in storing long-term memories is that lasting connectivity changes underlie the formation of representations of task-relevant parameters.

Duncan (2001) suggests that PFC neurons encode task parameters by regulating the activation of neurons that receive specific inputs. However, they do not speculate on whether the inputs to such PFC neurons are themselves strengthened or weakened in order modulate activity, and whether that modulation is maintained in altered synaptic weights or not. Hence, their adaptive coding model leaves open the role of PFC in long-term memory (see 1.2.3.2). On the other hand, Miller et al. (2002) propose that prefrontal cortex function in cognitive control and extracting task rules or contexts is implemented through specific patterns of sustained firing of individual neurons rather than through lasting synaptic changes. The authors name two reasons for their suggestion. First, cognitive control needs to span so many brain areas that top-down projections from PFC have to reach, that changing synaptic weights in all of them would require a large scale change. Second, changing the synaptic weights would likely be too slow for the level of behavioral flexibility that might be needed. However, the authors also suggest that, within PFC circuitry, a model of the task is built through repetition during learning, forming networks of associations. These associations could be established through changes in synaptic weights and connections.

In contrast, Fuster (2003) proposed that PFC fulfils a dual role. On the one hand, PFC temporally organizes goal-directed behavior, analogous to the previously discussed models (Duncan, 2001; Miller and Cohen, 2001; Miller et al., 2002). On the other hand, Fuster's theory argues that prefrontal cortex learns the relevant task rules or contexts and stores them as part of the long-term memory of a task. Network models of PFC function during category learning (Knoblich et al., 2002; Villagrasa et al., 2018), hypothesize Hebbian synaptic plasticity between areas ITC and PFC, i.e. direct connections from sensory areas, as the source for learned category representations in PFC neurons. This implies a role of PFC in the long-term memory of the learned task.

I observed that PFC neurons gradually acquire category selectivity reflecting the learning of a task rule (see 4.3.4), in line with the results from primate category learning (Antzoulatos and Miller, 2011; Villagrasa et al., 2018). A gradual acquisition of selectivity over multiple days points to a level of stability that could support memory of the learned categories or task rules. Studying semantic memory, it has not been tested whether such prefrontal representations stably encode parameters long-term and if their activation, specifically, is necessary to the learned behavior. After learning a task rule, broad optogenetic inactivation of PFC abolished the behavioral performance (Rikhye et al., 2018), but the study did not test whether neurons representing the task rule were crucial for the behavior rather than PFC as a whole. On the other hand, mouse prefrontal cortex has been implicated in the recall of episodic memory, i.e. mainly contextual fear memory (Frankland, 2004; Kitamura et al., 2017). These investigations have identified prefrontal cortex neurons that got recruited to encode fear memory. The activation of such 'engram cells' has been shown to have a causal role in the recall of remote, but not recent, memory. Importantly, an involvement in memory recall does not necessarily mean that a brain area also stores that memory (Simons and Spiers, 2003; Xiang and Brown, 2004).

In summary, studies so far have not been able to conclusively show whether, and how, prefrontal cortex is involved in storing long-term memories, independently from orchestrating goal-directed behavior. The gradual acquisition of category selectivity in mouse PFC indicates a more stable encoding of task-relevant parameters than suggested by adaptive coding models of

PFC (Duncan, 2001; Miller et al., 2002). This finding could point towards a role in long-term semantic memory. Future manipulations that are targeted to specific neuronal representation will be necessary to determine whether prefrontal cortex is part of the engram for semantic memory.

4.5 Outlook

The mouse model of category learning that I have established in this thesis opens up several possibilities for future work. Here, I will outline potential experiments that can contribute to elucidating neuronal underpinnings of category learning and computations in PFC that make it so relevant to goal-directed behavior.

Optogenetic manipulation (Boyden et al., 2005; Deisseroth, 2015) of prefrontal cortex activity in animals that are trained on a categorization task could be used to test the role of PFC in the memory of learned categories and in generalization to novel stimuli. Further, chronic optogenetic or chemogenetic (Armbruster et al., 2007; Roth, 2016) perturbation of PFC activity during the learning process could help to understand whether prefrontal networks take part in learning such a complex cognitive task. Potentially, neurons that are part of a characterized category representation could even be specifically targeted by conditionally expressing optogenetic tools under the control of an immediate early gene, like FosTRAP (Liu et al., 2012; Guenthner et al., 2013). Such a level of specificity would help to understand whether individual neuronal representations in PFC causally contribute to learned behavior or whether remaining population activity is sufficient. If individual neurons encoding learned categories are necessary for categorization behavior, that would support a role of PFC in long-term semantic memory.

In vivo whole brain activity mapping techniques, like functional ultrasound imaging (fUS; Macé et al., 2011), could be employed in mice that are performing the categorization task in order to search for brain regions that play a role in learning or performing the task. As an analogous to human fMRI, target areas for future investigations could be identified. We are currently starting an investigation using fUS, aiming to identify all brain regions that are involved in mouse category learning and at what learning stage they are involved. We hope to relate those results to network models of category learning ((see 1.2.4.3), especially Villagrasa et al., 2018) and, hereby, lay the groundwork to experimentally confirm the suggested interactions between brain areas.

The extensive knowledge of connectivity of the mouse brain (Oh et al., 2014; Abbott et al., 2020) and sophisticated tracing tools (Wickersham et al., 2007) can be used to test specific hypotheses about the circuitry underlying category learning. Transsynaptic tracing based on rabies virus allows for labeling of inputs to a brain area, for example inputs to prefrontal cortex. By expressing calcium indicators in such input populations, neuronal activity can be recorded from their axons in PFC. This enables investigating what type of information prefrontal cortex receives at what stage of learning and whether plastic changes happen at the level of inputs to PFC or rather within prefrontal cortex circuits. Understanding connections between brain areas with such detail could confirm predictions of category learning models, for example by identifying a striatal-based teaching signal (Villagrasa et al., 2018) or sensory input from visual areas (Knoblich et al., 2002). Further, by characterizing the types of inputs that PFC receives, one can potentially infer the computations that PFC neurons perform with those inputs, as has been done for retinothalamic connectivity (Rosón et al., 2019), and, hence, how mixed selectivity emerges in prefrontal cortex.

Here, I conclude that the mouse model of category learning has provided − and will continue to provide − valuable insights into this process of forming groups that is integral to our lives and by far not exclusive to us humans. The variety of tools that are available in mice will help to expand our understanding of neuronal computations that underlie category learning and will thereby complement the increasingly detailed investigations in humans.

Bibliography

Abbott, L. F., Bock, D. D., Callaway, E. M., Denk, W., Dulac, C., Fairhall, A. L., Fiete, I., Harris, K. M., Helmstaedter, M., Jain, V., Kasthuri, N., LeCun, Y., Lichtman, J. W., Littlewood, P. B., Luo, L., Maunsell, J. H., Reid, R. C., Rosen, B. R., Rubin, G. M., . . . Essen, D. C. V. (2020). The mind of a mouse. *Cell*, *182*(6), 1372–1376.

Abraham, N. M., Guerin, D., Bhaukaurally, K., & Carleton, A. (2012). Similar odor discrimination behavior in head-restrained and freely moving mice. *PLoS ONE*, *7*(12), e51789.

Ährlund-Richter, S., Xuan, Y., Lunteren, J. v., Kim, H., Ortiz, C., Dorocic, I., Meletis, K., & Carlén, M. (2019). A whole-brain atlas of monosynaptic input targeting four different cell types in the medial prefrontal cortex of the mouse. *Nature Neuroscience*, *22*(4), 657–668.

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, *12*(6), 977–987.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(1), 357–381.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3–19.

Andermann, M., Kerlin, A., & Reid, C. (2010). Chronic cellular imaging of mouse visual cortex during operant behavior and passive viewing. *Frontiers in Cellular Neuroscience*, *4*, 3.

Antzoulatos, E. G., & Miller, E. K. (2011). Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron*, *71*(2), 243–249.

Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S., & Roth, B. L. (2007). Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proceedings of the National Academy of Sciences*, *104*(12), 5163–5168.

Armstrong, C. M., DeVito, L. M., & Cleland, T. A. (2006). One-trial associative odor learning in neonatal mice. *Chemical Senses*, *31*(4), 343–349.

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, *21*(6), 1399–1407.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Crossley, M. J. (2010). Interactions between declarative and procedural-learning categorization systems. *Neurobiology of Learning and Memory*, *94*(1), 1–12.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632–656.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, T. W. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

Ashby, F. G., & Spiering, B. J. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 101–113.

Ashby, F., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*(5), 204–210.

Aust, U., & Huber, L. (2001). The role of item- and category-specific information in the discrimination of people versus nonpeople images by pigeons. *Animal Learning & Behavior*, *29*(2), 107–119.

Azizi, A. H., Pusch, R., Koenen, C., Klatt, S., Bröker, F., Thiele, S., Kellermann, J., Güntürkün, O., & Cheng, S. (2019). Emerging category representation in the visual forebrain hierarchy of pigeons (Columba livia). *Behavioural Brain Research*, *356*, 423–434.

Balaguer-Ballester, E., Lapish, C. C., Seamans, J. K., & Durstewitz, D. (2011). Attracting dynamics of frontal cortex ensembles during memory-guided decision-making. *PLoS Computational Biology*, *7*(5), e1002057.

Balsters, J. H., Zerbi, V., Sallet, J., Wenderoth, N., & Mars, R. B. (2020). Primate homologs of mouse cortico-striatal circuits. *eLife*, *9*, e53680.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, *437*(7062), 1158–1161.

Barretto, R. P. J., Messerschmidt, B., & Schnitzer, M. J. (2009). In vivo fluorescence imaging with high-resolution microlenses. *Nature Methods*, *6*(7), 511–512.

Behrmann, M., Geng, J. J., & Shomstein, S. (2004). Parietal cortex and attention. *Current Opinion in Neurobiology*, *14*(2), 212–217.

Benard, J., Stach, S., & Giurfa, M. (2006). Categorization of visual stimuli in the honeybee Apis mellifera. *Animal Cognition*, *9*(4), 257–270.

Bench, C., Frith, C., Grasby, P., Friston, K., Paulesu, E., Frackowiak, R., & Dolan, R. (1993). Investigations of the functional anatomy of attention using the stroop test. *Neuropsychologia*, *31*(9), 907–922.

Beninger, R. J. (1983). The role of dopamine in locomotor activity and learning. *Brain Research Reviews*, *6*(2), 173–196.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology*, *39*(1), 15–22.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, *81*(2), 179–209.

Biró, S., Lasztóczi, B., & Klausberger, T. (2019). A visual two-choice rule-switch task for head-fixed mice. *Frontiers in Behavioral Neuroscience*, *13*, 119.

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5–4 categories and the category advantage. *Memory & Cognition*, *31*(8), 1293–1301.

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614.

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263–1268.

Brincat, S. L., Siegel, M., Nicolai, C. v., & Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. *Proceedings of the National Academy of Sciences*, *115*(30), E7202–E7211.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. J. A. Barth.

Broschard, M. B., Kim, J., Love, B. C., & Freeman, J. H. (2020). Category learning in rodents using touchscreen-based tasks. *Genes, Brain and Behavior*, *20*(1), e12665.

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2021). Prelimbic cortex maintains attention to category-relevant information and flexibly updates category representations. *Neurobiology of Learning and Memory*, *185*, 107524.

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019b). Selective attention in rat visual category learning. *Learning & Memory*, *26*(3), 84–92.

Carlén, M. (2017). What constitutes the prefrontal cortex? *Science*, *358*(6362), 478–482.

Carli, M., Robbins, T., Evenden, J., & Everitt, B. (1983). Effects of lesions to ascending noradrenergic neurones on performance of a 5-choice serial reaction task in rats; implications for theories of dorsal noradrenergic bundle function based on selective attention and arousal. *Behavioural Brain Research*, *9*(3), 361–380.

Casale, M. B., Roeder, J. L., & Ashby, G. F. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449.

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., & Kim, D. S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, *499*(7458), 295–300.

Churchland, A. K., & Kiani, R. (2016). Three challenges for connecting model to mechanism in decision-making. *Current Opinion in Behavioral Sciences*, *11*, 74–80.

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B. J., Servan-Schreiber, D., & Noll, D. C. (1994). Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. *Human Brain Mapping*, *1*(4), 293–304.

Colacicco, G., Welzl, H., Lipp, H.-P., & Würbel, H. (2002). Attentional set-shifting in mice: modification of a rat paradigm, and evidence for strain-dependent variation. *Behavioural Brain Research*, *132*(1), 95–102.

Collins, D. P., Anastasiades, P. G., Marlin, J. J., & Carter, A. G. (2018). Reciprocal circuits linking the prefrontal cortex with dorsal and ventral thalamic nuclei. *Neuron*, *98*(2), 366–379.

Cools, A. R., Bercken, J. H. v. d., Horstink, M. W., Spaendonck, K. P. v., & Berger, H. J. (1984). Cognitive and motor shifting aptitude disorder in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *47*(5), 443–453.

Corbetta, M., Miezin, F., Dobmeyer, S., Shulman, G., & Petersen, S. (1990). Attentional modulation of neural processing of shape, color, and velocity in humans. *Science*, *248*(4962), 1556–1559.

Corkin, S. (1968). Acquisition of motor skill after bilateral medial temporal-lobe excision. *Neuropsychologia*, *6*(3), 255–265.

Costa, R. M., Cohen, D., & Nicolelis, M. A. (2004). Differential corticostriatal plasticity during fast and slow motor skill learning in mice. *Current Biology*, *14*(13), 1124–1134.

Creighton, S. D., Collett, H. A., Zonneveld, P. M., Pandit, R. A., Huff, A. E., Jardine, K. H., McNaughton, B. L., & Winters, B. D. (2019). Development of an "Object Category Recognition" task for mice: Involvement of muscarinic acetylcholine receptors. *Behavioral Neuroscience*, *133*(5), 527–536.

Cromer, J. A., Machon, M., & Miller, E. K. (2011). Rapid association learning in the primate prefrontal cortex in the absence of behavioral reversals. *Journal of Cognitive Neuroscience*, *23*(7), 1823–1828.

Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, *66*(5), 796–807.

David, S. V., Fritz, J. B., & Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences*, *109*(6), 2144–2149.

Deisseroth, K. (2015). Optogenetics: 10 years of microbial opsins in neuroscience. *Nature Neuroscience*, *18*(9), 1213–1225.

Denk, W., Strickler, J., & Webb, W. (1990). Two-photon laser scanning fluorescence microscopy. *Science*, *248*(4951), 73–76.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Devan, B., McDonald, R., & White, N. (1999). Effects of medial and lateral caudate-putamen lesions on place- and cue-guided behaviors in the water maze: relation to thigmotaxis. *Behavioural Brain Research*, *100*(1-2), 5–14.

Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Primate analogue of the Wisconsin Card Sorting Test: Effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behavioral Neuroscience*, *110*(5), 872–886.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Dolev, Y., & Nelson, X. J. (2014). Innate pattern recognition and categorization in a jumping spider. *PLoS ONE*, *9*(6), e97819.

Dräger, U. C. (1975). Receptive fields of single cells and topography in mouse visual cortex. *Journal of Comparative Neurology*, *160*(3), 269–289.

Duhamel, J. R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*(5040), 90–92.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829.

Eiselt, A.-K., Chen, S., Chen, J., Arnold, J., Kim, T., Pachitariu, M., & Sternson, S. M. (2021). Hunger or thirst state uncertainty is resolved by outcome evaluation in medial prefrontal cortex to guide decision-making. *Nature Neuroscience*, *24*(7), 907–912.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Estes, W. K. (1994). *Classification and Cognition*. Oxford University Press.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *26*(8), 3508–3526.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes. *Psychological Science*, *21*(3), 415–423.

Filoteo, J. V., Maddox, W. T., & Davis, J. D. (2001). Quantitative modeling of category learning in amnesic patients. *Journal of the International Neuropsychological Society*, *7*(1), 1–19.

Fitzgerald, J. K., Freedman, D. J., & Assad, J. A. (2011). Generalized associative representations in parietal cortex. *Nature Neuroscience*, *14*(8), 1075–1079.

Frankland, P. W. (2004). The involvement of the anterior cingulate cortex in remote contextual fear memory. *Science*, *304*(5672), 881–883.

Freedman, D. J., & Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, *443*(7107), 85–88.

Freedman, D. J., & Assad, J. A. (2016). Neuronal mechanisms of visual categorization: An abstract view on decision making. *Annual Review of Neuroscience*, *39*(1), 129–147.

Freedman, D. J., & Ibos, G. (2018). An integrative framework for sensory, motor, and cognitive functions of the posterior parietal cortex. *Neuron*, *97*(6), 1219–1234.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*(5502), 312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235–5246.

Furtak, S. C., Ahmed, O. J., & Burwell, R. D. (2012). Single neuron activity and theta modulation in postrhinal cortex during visual object discrimination. *Neuron*, *76*(5), 976–988.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654.

Fuster, J. M. (2002). Frontal lobe and cognitive development. *Journal of Neurocytology*, *31*, 373–385.

Fuster, J. M. (2003). Functional anatomy of the prefrontal cortex. In A. Beaumanoir, F. Andermann, P. Chauvel, L. Mira, & B. Zifkin (Eds.), *Frontal lobe seizures and epilepsies in children*. John Libbey Eurotext.

Fuster, J. M., Bodner, M., & Kroger, J. K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature*, *405*(6784), 347–351.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive psychology*, *25*(4), 524–575.

Gibson, K. R. (1991). Myelination and behavioral development: A comparative perspective on questions of neoteny, altriciality and intelligence. In A. C. P. Kathleen R. Gibson (Ed.), *Brain maturation and cognitive development: Comparative and cross-cultural perspectives.* Routledge.

Gifford III, G. W., MacLean, K. A., Hauser, M. D., & Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of Cognitive Neuroscience*, *17*(9), 1471–1482.

Giguere, M., & Goldman-Rakic, P. S. (1988). Mediodorsal nucleus: Areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. *Journal of Comparative Neurology*, *277*(2), 195–213.

Gilbert, C. D., & Wiesel, T. N. (1992). Receptive field dynamics in adult primary visual cortex. *Nature*, *356*(6365), 150–152.

Gluck, M. A., Oliver, L. M., & Myers, C. E. (1996). Late-training amnesic deficits in probabilistic category learning: a neurocomputational analysis. *Learning & Memory*, *3*(4), 326–340.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, *2*(1), 50–55.

Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.

Goldman-Rakic, P., & Schwartz, M. (1982). Interdigitation of contralateral and ipsilateral columnar projections to frontal association cortex in primates. *Science*, *216*(4547), 755–757.

Goldman-Rakic, P., Selemon, L., & Schwartz, M. (1984). Dual pathways connecting the dorsolateral prefrontal cortex with the hippocampal formation and parahippocampal cortex in the rhesus monkey. *Neuroscience*, *12*(3), 719–743.

Goltstein, P. M., Meijer, G. T., & Pennartz, C. M. (2018b). Conditioning sharpens the spatial representation of rewarded stimuli in mouse primary visual cortex. *eLife*, *7*, e37683.

Goltstein, P. M., Montijn, J. S., & Pennartz, C. M. (2015). Effects of isoflurane anesthesia on ensemble patterns of Ca2+ activity in mouse V1: Reduced direction selectivity independent of increased correlations in cellular activity. *PloS ONE*, *10*(2), e0118277.

Goltstein, P. M., Reinert, S., Bonhoeffer, T., & Hübener, M. (2021). Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *24*(10), 1441–1451.

Goltstein, P. M., Reinert, S., Glas, A., Bonhoeffer, T., & Hübener, M. (2018a). Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice. *PLoS ONE*, *13*(9), e0204066.

Göppert-Mayer, M. (1931). Über Elementarakte mit zwei Quantensprüngen. *Annalen der Physik*, *401*(3), 273–294.

Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology*, *15*(6), 638–644.

Guenthner, C. J., Miyamichi, K., Yang, H. H., Heller, H. C., & Luo, L. (2013). Permanent genetic access to transiently active neurons via TRAP: Targeted Recombination in Active Populations. *Neuron*, *78*(5), 773–784.

Güntürkün, O. (1997). Cognitive impairments after lesions of the neostriatum caudolaterale and its thalamic afferent in pigeons: Functional similarities to the mammalian prefrontal system? *Journal für Hirnforschung*, *38*(1), 133–143.

Güntürkün, O., Koenen, C., Iovine, F., Garland, A., & Pusch, R. (2018). The neuroscience of perceptual categorization in pigeons: A mechanistic hypothesis. *Learning & Behavior*, *46*(3), 229–241.

Guo, Z. V., Hires, S. A., Li, N., O'Connor, D. H., Komiyama, T., Ophir, E., Huber, D., Bonardi, C., Morandell, K., Gutnisky, D., Peron, S., Xu, N.-l., Cox, J., & Svoboda, K. (2014). Procedures for behavioral experiments in head-fixed mice. *PLoS ONE*, *9*(2), e88678.

Hamilton, L. W., & Flaherty, C. F. (1973). Interactive effects of deprivation in the albino rat. *Learning and Motivation*, *4*(2), 148–162.

Hampson, R. E., Pons, T. P., Stanford, T. R., & Deadwyler, S. A. (2004). Categorization in the monkey hippocampus: A possible mechanism for encoding information into memory. *Proceedings of the National Academy of Sciences*, *101*(9), 3184–3189.

Hampton, J. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language*, *34*(5), 686–708.

Han, X., Jing, M.-y., Zhao, T.-y., Wu, N., Song, R., & Li, J. (2017). Role of dopamine projections from ventral tegmental area to nucleus accumbens and medial prefrontal cortex in reinforcement behaviors assessed using optogenetic manipulation. *Metabolic Brain Disease*, *32*(5), 1491–1502.

Harvey, C. D., Coen, P., & Tank, D. W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, *484*(7392), 62–68.

Havenith, M. N., Zijderveld, P. M., Heukelum, S. v., Abghari, S., Glennon, J. C., & Tiesinga, P. (2018). The Virtual-Environment-Foraging task enables rapid training and single-trial metrics of attention in head-fixed mice. *Scientific Reports*, *8*(1), 17371.

Hayes, A. E., Davidson, M. C., Keele, S. W., & Rafal, R. D. (1998). Toward a functional analysis of the basal ganglia. *Journal of Cognitive Neuroscience*, *10*(2), 178–198.

Heaton, R. K., & Pendleton, M. G. (1981). Use of neuropsychological tests to predict adult patients' everyday functioning. *Journal of Consulting and Clinical Psychology*, *49*(6), 807–821.

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory.* Wiley.

Heisler, J. M., Morales, J., Donegan, J. J., Jett, J. D., Redus, L., & O'Connor, J. C. (2015). The attentional set shifting task: A measure of cognitive flexibility in mice. *Journal of Visualized Experiments*, (96), e51944.

Helmchen, F., & Denk, W. (2005). Deep tissue two-photon microscopy. *Nature Methods*, *2*(12), 932–940.

Herrnstein, R. J., & Loveland, D. H. (1964). Complex visual concept in the pigeon. *Science*, *146*(3643), 549–551.

Herrnstein, R. J., Loveland, D. H., & Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*(4), 285–302.

Hintiryan, H., Foster, N. N., Bowman, I., Bay, M., Song, M. Y., Gou, L., Yamashita, S., Bienkowski, M. S., Zingg, B., Zhu, M., Yang, X. W., Shih, J. C., Toga, A. W., & Dong, H.-W. (2016). The mouse cortico-striatal projectome. *Nature Neuroscience*, *19*(8), 1100–1114.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology*, *12*(4), 421–444.

Histed, M. H., Carvalho, L. A., & Maunsell, J. H. R. (2012). Psychophysical measurement of contrast sensitivity in the behaving mouse. *Journal of Neurophysiology*, *107*(3), 758–765.

Hoesen, G. W. V. (1982). The parahippocampal gyrus: New observations regarding its cortical connections in the monkey. *Trends in Neurosciences*, *5*, 345–350.

Hölscher, C., Schnee, A., Dahmen, H., Setia, L., & Mallot, H. (2005). Rats are able to navigate in virtual environments. *Journal of Experimental Biology*, *208*(3), 561–569.

Holtmaat, A., Bonhoeffer, T., Chow, D. K., Chuckowree, J., Paola, V. D., Hofer, S. B., Hübener, M., Keck, T., Knott, G., Lee, W.-C. A., Mostany, R., Mrsic-Flogel, T. D., Nedivi, E., Portera-Cailliau, C., Svoboda, K., Trachtenberg, J. T., & Wilbrecht, L. (2009). Long-term, high-resolution imaging in the mouse neocortex through a chronic cranial window. *Nature Protocols*, *4*(8), 1128–1144.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, *7*(6), 418–439.

Howard, J. D., Plailly, J., Grueschow, M., Haynes, J.-D., & Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nature Neuroscience*, *12*(7), 932–938.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106–154.

Hübener, M., & Bonhoeffer, T. (2010). Searching for engrams. *Neuron*, *67*(3), 363–371.

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological Monographs*, *28*(1), i–86.

Hwang, E., Willis, B. S., & Burwell, R. D. (2018). Prefrontal connections of the perirhinal and postrhinal cortices in the rat. *Behavioural Brain Research*, *354*, 8–21.

Jacobsen, C. (1928). Recent experiments on the function of the frontal lobes. *Psychological Bulletin*, *25*(1), 1–11.

Jacobsen, C. (1935). Functions of frontal association area in primates. *Archives of Neurology & Psychiatry*, *33*(3), 558–569.

Jaepel, J., Hübener, M., Bonhoeffer, T., & Rose, T. (2017). Lateral geniculate neurons projecting to primary visual cortex show ocular dominance plasticity in adult mice. *Nature Neuroscience*, *20*(12), 1708–1714.

James, W. (1913). *The principles of psychology*. Henry Holt; Company.

Jankovic, J., & Tolosa, E. (2007). *Parkinson's disease and movement disorders*. Lippincott Williams & Wilkins.

Jernigan, T. L., & Tallal, P. (1990). Late childhood changes in brain morphology observable with MRI. *Developmental Medicine & Child Neurology*, *32*(5), 379–385.

Jimenez, L. O., Tring, E., Trachtenberg, J. T., & Ringach, D. L. (2018). Local tuning biases in mouse primary visual cortex. *Journal of Neurophysiology*, *120*(1), 274–280.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, A. M. (1999). Building neural representations of habits. *Science*, *286*(5445), 1745–1749.

Jones, E. G., & Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*(4), 793–820.

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S., & Mintun, M. A. (1993). Spatial working memory in humans as revealed by PET. *Nature*, *363*(6430), 623–625.

Jourjine, N. (2017). Hunger and thirst interact to regulate ingestive behavior in flies and mammals. *BioEssays*, *39*(5), 1600261.

Jourjine, N., Mullaney, B. C., Mann, K., & Scott, K. (2016). Coupled sensing of hunger and thirst signals balances sugar and water consumption. *Cell*, *166*(4), 855–866.

Juczewski, K., Koussa, J. A., Kesner, A. J., Lee, J. O., & Lovinger, D. M. (2020). Stress and behavioral correlates in the head-fixed method: stress measurements, habituation dynamics, locomotion, and motor-skill learning in mice. *Scientific Reports*, *10*(1), 12245.

Jung, Y., Larsen, B., & Walther, D. B. (2018). Modality-independent coding of scene categories in prefrontal cortex. *Journal of Neuroscience*, *38*(26), 5969–5981.

Kase, D., Uta, D., Ishihara, H., & Imoto, K. (2015). Inhibitory synaptic transmission from the substantia nigra pars reticulata to the ventral medial thalamus in mice. *Neuroscience Research*, *97*, 26–35.

Katz, J. S., Wright, A. A., & Bachevalier, J. (2002). Mechanisms of same/different abstract-concept learning by rhesus monkeys (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(4), 358–368.

Keller, G. B., Bonhoeffer, T., & Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, *74*(5), 809–815.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309.

Kim, J., Castro, L., Wasserman, E. A., & Freeman, J. H. (2018). Dorsal hippocampus is necessary for visual categorization in rats. *Hippocampus*, *28*(6), 392–405.

Kim, T. H., Zhang, Y., Lecoq, J., Jung, J. C., Li, J., Zeng, H., Niell, C. M., & Schnitzer, M. J. (2016). Long-term optical access to an estimated one million neurons in the live mouse cortex. *Cell Reports*, *17*(12), 3385–3394.

Kirsch, J. A., Vlachos, I., Hausmann, M., Rose, J., Yim, M., Aertsen, A., & Güntürkün, O. (2009). Neuronal encoding of meaning: Establishing category-selective response patterns in the avian 'prefrontal cortex'. *Behavioural Brain Research*, *198*(1), 214–223.

Kitamura, T., Ogawa, S. K., Roy, D. S., Okuyama, T., Morrissey, M. D., Smith, L. M., Redondo, R. L., & Tonegawa, S. (2017). Engrams and circuits crucial for systems consolidation of a memory. *Science*, *356*(6333), 73–78.

Knoblich, U., Riesenhuber, M., Freedman, D. J., Miller, E. K., & Poggio, T. (2002). Visual categorization: How the monkey brain does it. In B. H, W. C, & P. T. Lee S-W (Eds.), *Biologically Motivated Computer Vision. BMCV 2002. Lecture Notes in Computer Science* (pp. 273–281). Springer, Berlin, Heidelberg.

Knopman, D., & Nissen, M. J. (1991). Procedural learning is impaired in Huntington's disease: Evidence from the serial reaction time task. *Neuropsychologia*, *29*(3), 245–254.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*(2), 106–120.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402.

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

Koenen, C., Pusch, R., Bröker, F., Thiele, S., & Güntürkün, O. (2016). Categories in the pigeon brain: A reverse engineering approach. *Journal of the Experimental Analysis of Behavior*, *105*(1), 111–122.

Komiyama, T., Sato, T. R., O'Connor, D. H., Zhang, Y.-X., Huber, D., Hooks, B. M., Gabitto, M., & Svoboda, K. (2010). Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature*, *464*(7292), 1182–1186.

Konishi, S., Kawazu, M., Uchida, I., Kikyo, H., Asakura, I., & Miyashita, Y. (1999). Contribution of working memory to transient activation in human inferior prefrontal cortex during performance of the Wisconsin Card Sorting Test. *Cerebral Cortex*, *9*(7), 745–753.

Kröner, S., & Güntürkün, O. (1999). Afferent and efferent connections of the caudolateral neostriatum in the pigeon (Columba livia): A retro- and anterograde pathway tracing study. *Journal of Comparative Neurology*, *407*(2), 228–260.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D., & Froemke, R. C. (2017). Parallel processing by cortical inhibition enables context-dependent behavior. *Nature Neuroscience*, *20*(1), 62–71.

Kudryavitskaya, E., Marom, E., Shani-Narkiss, H., Pash, D., & Mizrahi, A. (2021). Flexible categorization in the mouse olfactory bulb. *Current Biology*, *31*(8), 1616–1631.

Kuramoto, E., Pan, S., Furuta, T., Tanaka, Y. R., Iwai, H., Yamanaka, A., Ohno, S., Kaneko, T., Goto, T., & Hioki, H. (2016). Individual mediodorsal thalamic neurons project to multiple areas of the rat prefrontal cortex: A single neuron-tracing study using virus vectors. *Journal of Comparative Neurology*, *525*(1), 166–185.

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., Reddy, C. B., Kepecs, A., Harris, K. D., & Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron*, *105*(4), 700–711.

Lashley, K. S. (1950). *In search of the engram.* Cambridge University Press.

Ledergerber, D., Battistin, C., Blackstad, J. S., Gardner, R. J., Witter, M. P., Moser, M.-B., Roudi, Y., & Moser, E. I. (2021). Task-dependent mixed selectivity in the subiculum. *Cell Reports*, *35*(8), 109175.

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., Chen, L., Chen, L., Chen, T.-M., Chin, M. C., Chong, J., Crook, B. E., Czaplinska, A., Dang, C. N., Datta, S., … Jones, A. R. (2006). Genome-wide atlas of gene expression in the adult mouse brain. *445*(7124), 168–176.

Levy, D. J., & Glimcher, P. W. (2011). Comparing apples and oranges: Using reward-specific and reward-general subjective value representation in the brain. *Journal of Neuroscience*, *31*(41), 14693–14707.

Lim, S.-J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, *116*(10), 4671–4680.

Lindgren, H. S., Wickens, R., Tait, D. S., Brown, V. J., & Dunnett, S. B. (2013). Lesions of the dorsomedial striatum impair formation of attentional set in rats. *Neuropharmacology*, *71*, 148–153.

Lipp, H.-P., & Wolfer, D. P. (1998). Genetically modified mice and cognition. *Current Opinion in Neurobiology*, *8*(2), 272–280.

Little, D. M., Shin, S. S., Sisco, S. M., & Thulborn, K. R. (2006). Event-related fMRI of category learning: Differences in classification and feedback networks. *Brain and Cognition*, *60*(3), 244–252.

Liu, T., Slotnick, S. D., Serences, J. T., & Yantis, S. (2003). Cortical mechanisms of feature-based attentional control. *Cerebral Cortex*, *13*(12), 1334–1343.

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., & Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, *484*(7394), 381–385.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Low, R. J., Gu, Y., & Tank, D. W. (2014). Cellular resolution optical access to brain regions in fissures: Imaging medial prefrontal cortex and grid cells in entorhinal cortex. *Proceedings of the National Academy of Sciences*, *111*(52), 18739–18744.

Lu, M.-T., Preston, J. B., & Strick, P. L. (1994). Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. *Journal of Comparative Neurology*, *341*(3), 375–392.

Lui, J. H., Nguyen, N. D., Grutzner, S. M., Darmanis, S., Peixoto, D., Wagner, M. J., Allen, W. E., Kebschull, J. M., Richman, E. B., Ren, J., Newsome, W. T., Quake, S. R., & Luo, L. (2021). Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks. *Cell*, *184*(2), 489–506.

Macé, E., Montaldo, G., Cohen, I., Baulac, M., Fink, M., & Tanter, M. (2011). Functional ultrasound imaging of the brain. *Nature Methods*, *8*(8), 662–664.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Maddox, W. T., & Filoteo, J. V. (2001). Striatal contributions to category learning: Quantitative modeling of simple linear and complex nonlinear rule learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society*, *7*(6), 710–727.

Maddox, W. T., Filoteo, J. V., Lauritzen, J. S., Connally, E., & Hejl, K. D. (2005). Discontinuous categories affect information-integration but not rule-based category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 654–669.

Maddox, W., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*(3), 309–332.

Mank, M., Santos, A. F., Direnberger, S., Mrsic-Flogel, T. D., Hofer, S. B., Stein, V., Hendel, T., Reiff, D. F., Levelt, C., Borst, A., Bonhoeffer, T., Hübener, M., & Griesbeck, O. (2008). A genetically encoded calcium indicator for chronic in vivo two-photon imaging. *Nature Methods*, *5*(9), 805–811.

Mansouri, F. A. (2006). Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog. *Journal of Neuroscience*, *26*(10), 2745–2756.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Maunsell, J. H. R., & Cook, E. P. (2002). The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1063–1072.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McDonald, R. J., & White, N. M. (1994). Parallel information processing in the water maze: Evidence for independent memory systems involving dorsal striatum and hippocampus. *Behavioral and Neural Biology*, *61*(3), 260–270.

McDougall, W. (1923). Emotion, Instincts of Mammals and of Man: The Instinct of Combat. *Outline of Psychology* (pp. 139–142, 154–157). New York (Charles Scribner's Sons).

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241–253.

Merchant, H., Zainos, A., Hernández, A., Salinas, E., & Romo, R. (1997). Functional properties of primate putamen neurons during the categorization of tactile stimuli. *Journal of Neurophysiology*, *77*(3), 1132–1154.

Merre, P. L., Ährlund-Richter, S., & Carlén, M. (2021). The mouse prefrontal cortex: Unity in diversity. *Neuron*, *109*(12), 1925–1944.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1123–1136.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 90–100.

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of H.M. *Neuropsychologia*, *6*(3), 215–234.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category Size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799.

Miyawaki, A., Llopis, J., Heim, R., McCaffery, J. M., Adams, J. A., Ikura, M., & Tsien, R. Y. (1997). Fluorescent indicators for Ca2+based on green fluorescent proteins and calmodulin. *Nature*, *388*(6645), 882–887.

Monchi, O., Petrides, M., Strafella, A. P., Worsley, K. J., & Doyon, J. (2006). Functional role of the basal ganglia in the planning and execution of actions. *Annals of Neurology*, *59*(2), 257–264.

Morán, M. A., & Reinoso-Suárez, F. (1988). Topographical organization of the thalamic afferent connections to the motor cortex in the cat. *Journal of Comparative Neurology*, *270*(1), 64–85.

Morris, R. G. M., Garrud, P., Rawlins, J. N. P., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*(5868), 681–683.

Mountcastle, V. B., Lynch, J. C., Georgopoulos, A., Sakata, H., & Acuna, C. (1975). Posterior parietal association cortex of the monkey: Command functions for operations within extrapersonal space. *Journal of Neurophysiology*, *38*(4), 871–908.

Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*, *18*(6), 974–989.

Muller, R. U., & Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, *7*(7), 1951–1968.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686.

Nakajima, M., & Schmitt, L. I. (2019). Understanding the circuit basis of cognitive functions using mouse models. *Neuroscience Research*, *152*, 44–58.

Nelson, D. G. K. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*(6), 734–759.

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature*, *595*(7865), 80–84.

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Research*, *171*(2), 213–224.

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., Mesulam, M.-M., & Reber, P. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43.

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

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review*, *9*(1), 169–174.

Oh, S. W., Harris, J. A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A. M., Mortrud, M. T., Ouellette, B., Nguyen, T. N., Sorensen, S. A., Slaughterbeck, C. R., Wakeman, W., Li, Y., Feng, D., Ho, A., ... Zeng, H. (2014). A mesoscale connectome of the mouse brain. *Nature*, *508*(7495), 207–214.

Ohl, F. W., Scheich, H., & Freeman, W. J. (2001). Change in pattern of ongoing cortical activity with auditory category learning. *Nature*, *412*(6848), 733–736.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map.* Oxford: Clarendon Press.

Orban, G. A., Claeys, K., Nelissen, K., Smans, R., Sunaert, S., Todd, J. T., Wardak, C., Durand, J.-B., & Vanduffel, W. (2006). Mapping the parietal cortex of human and non-human primates. *Neuropsychologia*, *44*(13), 2647–2667.

Packard, M. G., & McGaugh, J. L. (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience*, *106*(3), 439–446.

Pan, W. X., Mao, T., & Dudman, J. T. (2010). Inputs to the dorsal striatum of the mouse reflect the parallel circuit architecture of the forebrain. *Frontiers in Neuroanatomy*, *4*, 147.

Pandya, D. N., & Yeterian, E. H. (1990). Prefrontal cortex in relation to other cortical areas in rhesus monkey: Architecture and connections. *Progress in Brain Research*, *85*, 63–94.

Pannunzi, M., Gigante, G., Mattia, M., Deco, G., Fusi, S., & Giudice, P. D. (2012). Learning selective top-down control enhances performance in a visual categorization task. *Journal of Neurophysiology*, *108*(11), 3124–3137.

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*, *20*(12), 1770–1779.

Pascual-Leone, A., Grafman, J., Clark, K., Stewart, M., Massaquoi, S., Lou, J.-S., & Hallett, M. (1993). Procedural learning in Parkinson's disease and cerebellar degeneration. *Annals of Neurology*, *34*(4), 594–602.

Pasupathy, A., & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, *433*(7028), 873–876.

Paul, E. J., & Ashby, F. G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*, *7*, 177.

Penfield, W., & Milner, B. (1958). Memory deficit produced by bilateral lesions in the hippocampal zone. *A.M.A. Archives of Neurology & Psychiatry*, *79*(5), 475–497.

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia*, *12*(3), 323–330.

Pfefferbaum, A., Mathalon, D. H., Sullivan, E. V., Rawles, J. M., Zipursky, R. B., & Lim, K. O. (1994). A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Archives of Neurology*, *51*(9), 874–887.

Phillips, D. P., & Irvine, D. R. (1981). Responses of single neurons in physiologically defined primary auditory cortex (AI) of the cat: frequency tuning and responses to intensity. *Journal of Neurophysiology*, *45*(1), 48–58.

Pinto, L., & Dan, Y. (2015). Cell-type-specific activity in prefrontal cortex during goal-directed behavior. *Neuron*, *87*(2), 437–450.

Pistell, P. J., Nelson, C. M., Miller, M. G., Spangler, E. L., Ingram, D. K., & Devan, B. D. (2009). Striatal lesions interfere with acquisition of a complex maze task in rats. *Behavioural Brain Research*, *197*(1), 138–143.

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrieli, J. D. E. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*(4), 564–574.

Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolic, I., Krupic, J., Bauza, M., Sahani, M., Keller, G. B., Mrsic-Flogel, T. D., & Hofer, S. B. (2015). Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron*, *86*(6), 1478–1490.

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*(1), 28–38.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*(1), 25–42.

Price, A., Filoteo, J. V., & Maddox, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*(5), 1213–1226.

Qi, X.-L., Meyer, T., Stanford, T. R., & Constantinidis, C. (2011). Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. *Cerebral Cortex*, *21*(12), 2722–2732.

Quintana, J., & Fuster, J. M. (1999). From perception to action: Temporal integrative functions of prefrontal and parietal neurons. *Cerebral Cortex*, *9*(3), 213–221.

Ragozzino, M. E. (2007). The contribution of the medial prefrontal cortex, orbitofrontal cortex, and dorsomedial striatum to behavioral flexibility. *Annals of the New York Academy of Sciences*, *1121*(1), 355–375.

Ragozzino, M. E., Detrick, S., & Kesner, R. P. (1999). Involvement of the prelimbic–infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. *Journal of Neuroscience*, *19*(11), 4585–4594.

Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, *393*(6685), 577–579.

Rainer, G., & Miller, E. K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, *27*(1), 179–189.

Rakic, P. (1988). Specification of cerebral cortical areas. *Science*, *241*(4862), 170–176.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*(5313), 821–824.

Rao, S. M., Bobholz, J. A., Hammeke, T. A., Rosen, A. C., Woodley, S. J., Cunningham, J. M., Cox, R. W., Stein, E. A., & Binder, J. R. (1997). Functional MRI evidence for subcortical participation in conceptual reasoning skills. *NeuroReport*, *8*(8), 1987–1993.

Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, *17*(12), 1784–1792.

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences*, *95*(2), 747–750.

Reed, J. M., & Squire, L. R. (1999). Impaired transverse patterning in human amnesia is a special case of impaired memory for two-choice discrimination tasks. *Behavioral Neuroscience*, *113*(1), 3–9.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Reilly, D. L., Cooper, L. N., & Elbaum, C. (1982). A neural model for category learning. *Biological Cybernetics*, *45*(1), 35–41.

Reinert, S., Hübener, M., Bonhoeffer, T., & Goltstein, P. M. (2021). Mouse prefrontal cortex represents learned rules for categorization. *Nature*, *593*(7859), 411–417.

Remington, E. D., Osmanski, M. S., & Wang, X. (2012). An operant conditioning method for studying auditory behaviors in marmoset monkeys. *PLoS ONE*, *7*(10), e47895.

Rickard, T. C., & Grafman, J. (1998). Losing their configural mind: Amnesic patients fail on transverse patterning. *Journal of Cognitive Neuroscience*, *10*(4), 509–524.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590.

Rikhye, R. V., Gilra, A., & Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, *21*(12), 1753–1763.

Ritter, R. C., & Epstein, A. N. (1974). Saliva lost by grooming: A major item in the rat's water economy. *Behavioral Biology*, *11*(4), 581–585.

Robbins, T. (2002). The 5-choice serial reaction time task: Behavioural pharmacology and functional neurochemistry. *Psychopharmacology*, *163*(3-4), 362–380.

Robinson, A. L., Heaton, R. K., Lehman, R. A., & Stilson, D. W. (1980). The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. *Journal of Consulting and Clinical Psychology*, *48*(5), 605–614.

Rolls, E. T. (2000). Memory systems in the brain. *Annual Review of Psychology*, *51*(1), 599–630.

Romani, G., Williamson, S., & Kaufman, L. (1982). Tonotopic organization of the human auditory cortex. *Science*, *216*(4552), 1339–1340.

Rose, J. E., & Woolsey, C. N. (1948). The orbitofrontal cortex and its connections with the mediodorsal nucleus in rabbit, sheep and cat. *Research publications - Association for Research in Nervous and Mental Disease*, *27*(1), 210–232.

Rose, T., Goltstein, P. M., Portugues, R., & Griesbeck, O. (2014). Putting a finishing touch on GECIs. *Frontiers in Molecular Neuroscience*, *7*, 88.

Rosedahl, L. A., Eckstein, M. P., & Ashby, F. G. (2018). Retinal-specific category learning. *Nature Human Behaviour*, *2*(7), 500–506.

Rosón, M. R., Bauer, Y., Kotkat, A. H., Berens, P., Euler, T., & Busse, L. (2019). Mouse dLGN receives functional input from a diverse population of retinal ganglion cells with limited convergence. *Neuron*, *102*(2), 462–476.

Rossi, A. F., Bichot, N. P., Desimone, R., & Ungerleider, L. G. (2007). Top–Down attentional deficits in macaques with lesions of lateral prefrontal cortex. *Journal of Neuroscience*, *27*(42), 11306–11314.

Rossi-Pool, R., Salinas, E., Zainos, A., Alvarez, M., Vergara, J., Parga, N., & Romo, R. (2016). Emergence of an abstract categorical code enabling the discrimination of temporally structured tactile stimuli. *Proceedings of the National Academy of Sciences*, *113*(49), E7966–E7975.

Roth, B. L. (2016). DREADDs for neuroscientists. *Neuron*, *89*(4), 683–694.

Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal cortex activity during flexible categorization. *Journal of Neuroscience*, *30*(25), 8519–8528.

Rugg, M. D., Otten, L. J., & Henson, R. N. A. (2002). The neural basis of episodic memory: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1097–1110.

Runyan, C. A., Piasini, E., Panzeri, S., & Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature*, *548*(7665), 92–96.

Saleem, A. B., Ayaz, A., Jeffery, K. J., Harris, K. D., & Carandini, M. (2013). Integration of visual motion and locomotion in mouse visual cortex. *Nature Neuroscience*, *16*(12), 1864–1869.

Sanders, J. I., & Kepecs, A. (2012). Choice ball: A response interface for two-choice psychometric discrimination in head-fixed mice. *Journal of Neurophysiology*, *108*(12), 3416–3423.

Santiago, A. N., Makowicz, E. A., Du, M., & Aoki, C. (2021). Food restriction engages prefrontal corticostriatal cells and local microcircuitry to drive the decision to run versus conserve energy. *Cerebral Cortex*, *31*(6), 2868–2885.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473–500.

Schwarz, C., Hentschke, H., Butovas, S., Haiss, F., Stüttgen, M. C., Gerdjikov, T. V., Bergner, C. G., & Waiblinger, C. (2010). The head-fixed behaving rat—Procedures and pitfalls. *Somatosensory & Motor Research*, *27*(4), 131–148.

Seamans, J. K., Lapish, C. C., & Durstewitz, D. (2008). Comparing the prefrontal cortex of rats and primates: Insights from electrophysiology. *Neurotoxicity Research*, *14*(2-3), 249–262.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265–278.

Seger, C. A., & Cincotta, C. M. (2002). Striatal activity in concept learning. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 149–161.

Seger, C. A., Dennison, C. S., Lopez-Paniagua, D., Peterson, E. J., & Roark, A. A. (2011). Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. *NeuroImage*, *55*(4), 1739–1753.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*(1), 203–219.

Semendeferi, K., Lu, A., Schenker, N., & Damasio, H. (2002). Humans and great apes share a large frontal cortex. *Nature Neuroscience*, *5*(3), 272–276.

Seyfarth, R., Cheney, D., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803.

Seyfarth, R. M., & Cheney, D. L. (1986). Vocal development in vervet monkeys. *Animal Behaviour*, *34*(6), 1640–1658.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6), 509–523.

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94–102.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Shimamura, A. P., Janowsky, J. S., & Squire, L. R. (1990). Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. *Neuropsychologia*, *28*(8), 803–813.

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience*, *24*(47), 10702–10706.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320.

Simons, J. S., & Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience*, *4*(8), 637–648.

Sinex, D. G., Burdette, L. J., & Pearlman, A. L. (1979). A psychophysical investigation of spatial vision in the normal and reeler mutant mouse. *Vision Research*, *19*(8), 853–857.

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *Journal of General Psychology*, *12*(1), 66–77.

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18*(2), 414–421.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (Macaca mulatta) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(1), 54–65.

Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., Spiering, B., Beran, M. J., Church, B. A., Ashby, F. G., & Grace, R. C. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2355–2369.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 659–680.

Smith, J. D., Zakrzewski, A. C., Johnston, J. J. R., Roeder, J. L., Boomer, J., Ashby, F. G., & Church, B. A. (2015). Generalization of category knowledge and dimensional categorization in humans (Homo sapiens) and nonhuman primates (Macaca mulatta). *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(4), 322–335.

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2021). The geometry of concept learning. *bioRxiv*.

Sowell, E. R., Thompson, P. M., Holmes, C. J., Jernigan, T. L., & Toga, A. W. (1999). In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience*, *2*(10), 859–861.

Spellman, T., Svei, M., Kaminsky, J., Manzano-Nieves, G., & Liston, C. (2021). Prefrontal deep projection neurons enable cognitive flexibility via persistent feedback monitoring. *Cell*, *184*(10), 2750–2766.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*(5026), 1380–1386.

Squire, L. R. (1987). *Memory and brain*. Oxford University Press.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology*, *5*(2), 169–177.

Staddon, J. E. R. (2016). *Adaptive behavior and learning*. Cambridge University Press.

Stosiek, C., Garaschuk, O., Holthoff, K., & Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, *100*(12), 7319–7324.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Svoboda, K., & Yasuda, R. (2006). Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron*, *50*(6), 823–839.

Swaminathan, S. K., & Freedman, D. J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nature Neuroscience*, *15*(2), 315–320.

Thomas, K. R., & Capecchi, M. R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, *51*(3), 503–512.

Thorndike, E. L. (1913). *Educational psychology, Vol 1: The original nature of man*. Teachers College.

Tian, L., Hires, S. A., Mao, T., Huber, D., Chiappe, M. E., Chalasani, S. H., Petreanu, L., Akerboom, J., McKinney, S. A., Schreiter, E. R., Bargmann, C. I., Jayaraman, V., Svoboda, K., & Looger, L. L. (2009). Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods*, *6*(12), 875–881.

Toth, L. A., & Gardiner, T. W. (2000). Food and water restriction protocols: Physiological and behavioral considerations. *Journal of the American Association for Laboratory Animal Science*, *39*(6), 9–17.

Tucci, V., Hardy, A., & Nolan, P. M. (2006). A comparison of physiological and behavioural parameters in C57BL/6J mice undergoing food or water restriction regimes. *Behavioural Brain Research*, *173*(1), 22–29.

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, *40*(4), 385–398.

Tulving, E., Markowitsch, H. J., Craik, F. I. M., Habib, R., & Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, *6*(1), 71–79.

Uylings, H. B., & Eden, C. G. v. (1991). Chapter 3 Qualitative and quantitative comparison of the prefrontal cortex in rat and in primates, including humans. *Progress in Brain Research*, *85*, 31–62.

Uylings, H. B., Groenewegen, H. J., & Kolb, B. (2003). Do rats have a prefrontal cortex? *Behavioural Brain Research*, *146*(1-2), 3–17.

Vendrell, P., Junqué, C., Pujol, J., Jurado, M., Molet, J., & Grafman, J. (1995). The role of prefrontal regions in the Stroop task. *Neuropsychologia*, *33*(3), 341–352.

Vermaercke, B., Cop, E., Willems, S., D'Hooge, R., & Beeck, H. P. d. (2014). More complex brains are not always better: Rats outperform humans in implicit category-based generalization by implementing a similarity-based strategy. *Psychonomic Bulletin & Review*, *21*(4), 1080–1086.

Villagrasa, F., Baladron, J., Vitay, J., Schroll, H., Antzoulatos, E. G., Miller, E. K., & Hamker, F. H. (2018). On the role of cortex-basal ganglia interactions for category learning: A neuro-computational approach. *Journal of Neuroscience*, *38*(44), 9551–9562.

Vogels, R., Sary, G., Dupont, P., & Orban, G. A. (2002). Human brain regions involved in visual categorization. *NeuroImage*, *16*(2), 401–414.

Volz, H.-P., Gaser, C., Häger, F., Rzanny, R., Mentzel, H.-J., Kreitschmann-Andermahr, I., Kaiser, W. A., & Sauer, H. (1997). Brain activation during cognitive stimulation with the Wisconsin Card Sorting Test – a functional MRI study on healthy volunteers and schizophrenics. *Psychiatry Research: Neuroimaging*, *75*(3), 145–157.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*(1), 168–176.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953–956.

Wang, Q., Sporns, O., & Burkhalter, A. (2012). Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. *Journal of Neuroscience*, *32*(13), 4386–4399.

Warrington, E. K., & Weiskrantz, L. (1968). New method of testing long-term retention with special reference to amnesic patients. *Nature*, *217*(5132), 972–974.

Watanabe, K., & Funahashi, S. (2015). A dual-task paradigm for behavioral and neurobiological studies in nonhuman primates. *Journal of Neuroscience Methods*, *246*, 1–12.

Watanabe, S. (2013). Preference for and discrimination of paintings by mice. *PLoS ONE*, *8*(6), e65335.

Watanabe, S. (2017). Paintings discrimination by mice: Different strategies for different paintings. *Behavioural Processes*, *142*, 126–130.

Watanabe, S., Sakamoto, J., & Wakita, M. (1995). Pigeons' discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior*, *63*(2), 165–174.

Welzl, H., D'Adamo, P., & Lipp, H.-P. (2001). Conditioned taste aversion as a learning and memory paradigm. *Behavioural Brain Research*, *125*(1-2), 205–213.

Whishaw, I. Q., Mittleman, G., Bunch, S., & Dunnett, S. B. (1987). Impairments in the acquisition, retention and selection of spatial navigation strategies after medial caudate-putamen lesions in rats. *Behavioural Brain Research*, *24*(2), 125–138.

Whitlock, J. R. (2014). Navigating actions through the rodent parietal cortex. *Frontiers in Human Neuroscience*, *8*, 293.

Whitlock, J. R. (2017). Posterior parietal cortex. *Current Biology*, *27*(14), R691–R695.

Wickens, J. R., Alexander, M. E., & Miller, R. (1991). Two dynamic modes of striatal function under dopaminergic-cholinergic control: Simulation and analysis of a model. *Synapse*, *8*(1), 1–12.

Wickersham, I. R., Lyon, D. C., Barnard, R. J., Mori, T., Finke, S., Conzelmann, K.-K., Young, J. A., & Callaway, E. M. (2007). Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron*, *53*(5), 639–647.

Winocur, G., & Eskes, G. (1998). Prefrontal cortex and caudate nucleus in conditional associative learning: Dissociated effects of selective brain lesions in rats. *Behavioral Neuroscience*, *112*(1), 89–101.

Wise, R. A., & Rompre, P. P. (1989). Brain dopamine and reward. *Annual Review of Psychology*, *40*(1), 191–225.

Wyttenbach, R. A., May, M. L., & Hoy, R. R. (1996). Categorical perception of sound frequency by crickets. *Science*, *273*(5281), 1542–1544.

Xiang, J.-Z., & Brown, M. W. (2004). Neuronal responses related to long-term recognition memory processes in prefrontal cortex. *Neuron*, *42*(5), 817–829.

Xin, Y., Zhong, L., Zhang, Y., Zhou, T., Pan, J., & Xu, N.-l. (2019). Sensory-to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. *Neuron*, *103*(5), 909–921.

Yamazaki, Y., Aust, U., Huber, L., Hausmann, M., & Güntürkün, O. (2007). Lateralized cognition: Asymmetrical and complementary strategies of pigeons during discrimination of the "human concept". *Cognition*, *104*(2), 315–344.

Yantis, S., Schwarzbach, J., Serences, J. T., Carlson, R. L., Steinmetz, M. A., Pekar, J. J., & Courtney, S. M. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience*, *5*(10), 995–1002.

Yantis, S., & Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology*, *13*(2), 187–193.

Ye, L., Allen, W. E., Thompson, K. R., Tian, Q., Hsueh, B., Ramakrishnan, C., Wang, A.-C. C., Jennings, J. H., Adhikari, A., Halpern, C. H., Witten, I. B., Barth, A. L., Luo, L., McNab, J. A., & Deisseroth, K. (2016). Wiring and molecular features of prefrontal ensembles representing distinct experiences. *Cell*, *165*(7), 1776–1788.

Yuste, R., & Denk, W. (1995). Dendritic spines as basic functional units of neuronal integration. *Nature*, *375*(6533), 682–684.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(2), 387–398.

Zempeltzi, M. M., Kisse, M., Brunk, M. G. K., Glemser, C., Aksit, S., Deane, K. E., Maurya, S., Schneider, L., Ohl, F. W., Deliano, M., & Happel, M. F. K. (2020). Task rule and choice are reflected by layer-specific processing in rodent auditory cortical microcircuits. *Communications Biology*, *3*(1), 345.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K. R., & Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition & Behavior Reviews*, *3*, 13–45.

Zheng, M., Xu, J., Keniston, L., Wu, J., Chang, S., & Yu, L. (2021). Choice-dependent cross-modal interaction in the medial prefrontal cortex of rats. *Molecular Brain*, *14*(1), 13.

Zhong, L., Zhang, Y., Duan, C. A., Deng, J., Pan, J., & Xu, N.-l. (2019). Causal contributions of parietal cortex to perceptual decision-making during stimulus categorization. *Nature Neuroscience*, *22*(6), 963–973.

Zola-Morgan, S., Cohen, N. J., & Squire, L. R. (1983). Recall of remote episodic memory in amnesia. *Neuropsychologia*, *21*(5), 487–500.

Acknowledgements

I first and foremost want to thank Pieter. Without you, this thesis, these projects and those studies would not be. From circling cells, via listening to daft punk during surgeries, to trying to become a mouse-whisperer, to our Sunday-Funday's in the lab, I think we both can say it's been an adventure. And wow, I would not have believed, where we got, if you'd told me in 2014. Thank you for guiding my first baby-steps in neuroscience, for putting enough trust in me to take me on as your student, for all the countless hours in lab, for the patience with all my "Sorry, quick question: . . . "s, and most importantly, for showing me the scientist I want to be. I am grateful for all the things I have learned from you, be it the Illustrator-tip-of-the-day, how to solder properly or even how to jump-start a car. Thank you for being the world's best supervisor and a wonderful friend!

Mark, I want to thank you for welcoming me to the lab a long time ago and all the support I've gotten from you in all those years. I am very grateful that you agreed to be my Doktorvater, helped me get into the GSN, together with Pieter came up with a wonderful project for me and guided that project with valuable insight.

I am extremely grateful to you, Tobias, that you accepted me the first time I applied to your lab, fresh out of high school almost 10 years ago. And then again, as a HiWi and undergrad student and then again as a PhD student. Thank you for letting me be part of that wonderful team that you built, for recognizing the potential of my project and for giving the support that helped see it through.

Christian, thank you for agreeing to co-supervise my PhD project, for admitting me to the RTG, a group of excellent people, that I was proud to be a part of, and for all I have learned with you on my TAC over the years.

Volker, I don't think anyone has shaped my early path into science as much as you have. Thank you so much for taking me on as an intern, when I was too nervous to do basic maths in my head. Thank you for showing me how fun science can be and what commitment to science means. Thank you for answering 12.000 dumb questions, not running away from a 'Voooolker. . . ?' and always being just one jump away from the thing that I needed to keep an experiment running, be it a telephone cable or good advice.

Claudia, Frank and Max, thank you so much for all your work that made my projects possible! Be it a perfect immunostaining, exactly the right mouse at the right time or the countless times I needed help with a setup, labview or a matlab license, I could rely on you to be there and happy to help.

Special gratitude goes to Meike and Verena for their endless support with all the admin work, for keeping the lab (Meike) and the RTG/GSN (Verena) running and, most importantly, for your friendship. Meike, I am so happy you joined the lab and thankful for the breaks from science, be it with coffee or acro-yoga. Verena, I don't think I could have navigated PhD life, fulfilling all grad-school requirements all while doing actual science, without you. Also, you shaped all the wonderful experiences with the RTG, especially an unforgettable trip to Venice. Thank you!

I thank the Graduate School of Systemic Neurosciences and the RTG 2175 and their coordination teams for the support during my PhD, the training, and for providing a network of brilliant scientists and a fun environment.

A heartfelt thanks goes to all lab members, the Bon-Bros, for being a wonderful team and for all the experiences we shared. The everyday lunches and coffees, the retreats, the holidays and our spectacular happy hours. Especially, thank you Isa, David and Hiranmay, climbing buddies, former almost-office buddies, 2016 crew, TST, you made my time in the lab so much fun!

Lena, a special thanks goes to you, for so much outside-of-lab support, for always having an open ear and for our Sausalitos evenings that are a tradition by now!

Lastly, and most importantly, I want to thank my family. You all, and especially you, Mama, have guided me, helped me find my way into completely unchartered territory and have supported me, regardless of whether I needed lifting up or cheering on. Joël, my office mate, best friend and partner all at once: words can't describe how grateful I am to have you with me. I don't know what I would have done without you, our countless brainstorming sessions and the super valuable insights and ideas I got from you. I cannot thank you enough for your unconditional support, scientifically and emotionally. I can only say, you made my time in the lab the incredible experience it was and I am looking forward to our next adventure!

Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation *One rule to rule them all: Representations of learned rules for categorization in mouse prefrontal cortex* selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation *One rule to rule them all: Representations of learned rules for categorization in mouse prefrontal cortex* is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Sandra Reinert
München, den 29.10.21
Munich, 29.10.21

List of publications

Publications included in this thesis

Goltstein, P. M.*, **Reinert, S.***, Glas, A.*, Bonhoeffer, T., Hübener, M. (2018). Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice. *PLoS ONE, 13*(9), e0204066.
Authors with a * contributed equally.

Reinert, S., Hübener, M., Bonhoeffer, T., Goltstein, P. M. (2021). Mouse prefrontal cortex represents learned rules for categorization. *Nature, 593*(7859), 411–417.

Publications not included in this thesis

Goltstein, P. M., **Reinert, S.**, Bonhoeffer, T., Hübener, M. (2021). Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nature Neuroscience, 24*(10), 1441–1451.

Declaration of author contributions

Chapter 2

Goltstein, P.M.,* ***Reinert, S.****, *Glas, A.*, Bonhoeffer, T., Hübener, M. (2018) Food and water restriction lead to differential learning behaviors in a head-fixed two-choice visual discrimination task for mice. PLoS ONE 13(9): e0204066.*
Authors with a * contributed equally.

All listed authors contributed to the study: Pieter M. Goltstein, Sandra Reinert, Annet Glas, Tobias Bonhoeffer, and Mark Hübener. The study was designed by all authors. Data for this study was acquired by Pieter M. Goltstein, Sandra Reinert, Annet Glas and Mark Hübener. The data was analyzed by Pieter M. Goltstein. The manuscript was written and revised by all authors. All authors approved the final version of the manuscript.

Chapter 3

Reinert, S., *Hübener, M., Bonhoeffer, T., Goltstein, P.M. (2021) Mouse prefrontal cortex represents learned rules for categorization. Nature 593, 411–417.*

All listed authors contributed to the study: Sandra Reinert, Mark Hübener, Tobias Bonhoeffer and Pieter M. Goltstein. The study was designed by all authors. Data for this study was acquired by Sandra Reinert. Sandra Reinert and Pieter M. Goltstein analyzed the data. All authors wrote and revised the manuscript and approved the final version of the manuscript.

Prof. Dr. Tobias Bonhoeffer

Prof. Dr. Mark Hübener

Dr. Pieter Goltstein

Dr. Annet Glas

Sandra Reinert