The primacy of behavioral research for understanding the brain

Yael Niv[1]

[1]Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA.

Abstract: Understanding the brain requires us to answer both what the brain does, and how it does it. Using a series of examples, I make the case that behavior is often more useful than neuroscientific measurements for answering the first question[1]. Moreover, I show that even for "how" questions that pertain to neural mechanism, a well-crafted behavioral paradigm can offer deeper insight and stronger constraints on computational and mechanistic models than do many highly challenging (and very expensive) neural studies. I conclude that purely behavioral research is essential for understanding the brain—especially its cognitive functions—contrary to the opinion of prominent funding bodies and some scientific journals, who erroneously place neural data on a pedestal and consider behavior to be subsidiary.

Introduction

In an era of increasingly more precise methods for measuring and perturbing neurons in the brain, it often seems that with more neural data, we will soon make untold breakthroughs in understanding the brain. Such anticipation has heralded neuroscience-data-focused projects such as the Brain Initiative and the Human Connectome Project. The focus on neural measurement has been accompanied by the demotion of animal and human behavioral research—the mainstays of psychology from where understanding of the brain and mind originally hailed—to being "only" behavior, that is, insufficient, and even irrelevant to the neuroscientific quest. This widespread neuroscience chauvinism even infiltrates the way we think about and discuss our findings: decision making is often described as executed by neurons, not by animals, and whole experiments in awake behaving animals are sometimes described from the point of view of neurons, not the animal housing them: 'neurons were presented with images...'

I believe that this approach is fundamentally flawed. Assuming that a process such as decision making can be understood solely by looking at single neurons, or even their ensembles, is like attempting to understand why people in Australia drive on the left side of the road from examination of their DNA. Neural firing patterns are the wrong level for investigating many pressing questions in neuroscience. Even if we could

[1]My arguments here have also been published, in similar form, as a chapter in Lerner, Cullen, & Leslie, 2020

measure all the neurons in the brain with arbitrary precision, without an incisive *behavioral* paradigm we would not be able to answer many neuroscientific questions. Indeed, the insights about the brain that I have gleaned through my own research have almost all come from behavioral data, which explains the progressive decrease of neuroimaging experiments that we choose to run.[2]

An extremely sobering reminder of the fact that a full neural description does not guarantee understanding of what the brain does or how it does it, is the nematode *Caenorhabditis elegans,* whose behavior we are still far from being able to predict despite it having only 302 neurons and probably the best characterized nervous system in the universe (Bargmann & Marder, 2013). The fact that full knowledge of a circuit does not equate with understanding of that circuit suggests that when trying to understand the brain, and in particular high-order cognitive functions, focusing on neurons to the exclusion of other levels of inquiry, and without careful thought about behavioral paradigms that allow us to meaningfully interrogate the neural substrate, might lead us down an expensive and less-than-revealing rabbit hole. For the sake of better understanding the brain, we should therefore reverse the current "hierarchy" (in which neural measurements are seen as basic and fundamental, and behavior is an optional component that cannot stand on its own) and restore behavioral research to its historical place of primacy and necessity.

One might argue that the reason the majority of discoveries that have withstood time are due to behavioral research is that until recently we did not have access for incisive tools for neuroscientific research—non-invasive magnetic resonance imaging of humans and optogenetic manipulations are fairly new techniques. However, cutting-edge methods for recording and manipulating the brain have, for the most part, verified what we already knew from behavior, rather than led to truly novel insight. Of course, neuroscientific measurements (or brain perturbation studies) are necessary for answering questions about localization—*where* in the brain different (behaviorally-identified) cognitive functions are implemented. But even today, behavior remains the standard by which theories about the *roles* and *computations* carried out by different neural structures are tested (even for theories that do not seek to model behavior directly, see for example, Wang, 2008). If we ask ourselves candidly "what has neuroscience taught us about cognition that we did not already know from behavior?" we realize that unfortunately the answer is "not very much." Therefore, in the name of expediency, we have a moral obligation to our subjects of study, our taxpayer funders, and the patients waiting for cures, to reverse the trend that suggests that we should dispense with behavioral research and focus only on neural mechanisms.

[2]This is not because functional neuroimaging (fMRI) experiments are useless. FMRI data, like any other data, be it choice behavior, reaction times, eye movements, can answer some questions and not others. My research focuses on understanding what algorithms the brain uses for decision making, and how these are affected by mental illness (i.e., computational- and algorithmic-level questions, rather than implementation-level questions, according to Marr, 1982). For these particular questions, I often find behavioral data together with computational models to provide more constraints than do neural data, and am hard-pressed to come up with questions for which the neural data will tell us something beyond the behavior. However, there are exceptions, and I have used neural data, in conjunction with specifically-tailored behavioral experiments and computational models, to formally compare between computational models (admittedly, arriving at the same conclusion as was seen in the behavioral data) (Niv et al., 2015), and to contrast computational algorithms that could not be differentiated behaviorally (Niv, Edlund, Dayan, & O'Doherty, 2012). Different from Coltheart (2006), I am not arguing here that fMRI—or any neuroscience experiment—is useless, only that without incisive behavioral paradigms, neuroscience is not nearly sufficient for achieving the understanding of cognition that we aim for.

What has behavior taught us about the brain?

Long before the advent of neuroscience methods such as optogenetics and DREADDs (Designer Receptors Exclusively Activated by Designer Drugs), ingenious behavioral experiments made great strides in identifying the latent cognitive processes that underlie behavior—the ultimate output of the brain. For instance, even in low-level perception, arguably the area of neuroscience farthest from behavioral output, scientists were able to infer that color vision is due to three types of neurons (retinal cones), and were even able to estimate the cones' wavelength sensitivity from psychophysics experiments of color matching and chromatic adaptation and studies in color-blind observers (e.g. Stiles, 1959; Stiles & Burch, 1959). This occurred decades before physiologists were able to patch and measure the wavelength response properties of individual cones and verify the original behavior-based predictions (Baylor, Nunn, & Schnapf, 1987).

At the other end of the spectrum, the elusive domains of high-level cognition and cognitive control, ideas about the role and information content of attentional signals feeding back into low-level perceptual processing areas were also derived from simple, but revealing, behavioral paradigms such as visual search and pop-out (Hochstein & Ahissar, 2002), and from experiments delineating the extent to which practice effects generalize across simple visual tasks (Ahissar & Hochstein, 1993, 1997). These two examples from the domain of vision are representative of others in all levels of research into the workings of the brain and the mind. In this section, I describe in detail a small selection of results from behavioral studies that illuminate brain processes, each going beyond what would have been achievable even using unrealistic, whole-brain, single-cell resolution neural measurements.

I will start with my personal favorite: a rat is trained to run in a T-shaped maze, from the base of the T to the right arm, in order to obtain food. At the critical junction, the rat may be choosing to go right due to internal, egocentric cues (i.e., turning right relative to its own body), or based on external, allocentric cues (turning towards a certain location in space, for example to the side of the room that has a window). Imagine you could record activity from anywhere in the rat's brain during this experiment. How could you determine which of the two strategies the rat was using? Even knowing what we know about navigation, spatial maps in the hippocampus, and head direction cells, it is not clear what you should look for in your recordings. In particular, the fact that the external cues of the window and blinds are represented in some brain area does not mean the rat is using these cues to guide its actions. Arguably, even if you had recordings of every neuron in the brain during this task, it would be extremely difficult to answer this simple question regarding the rat's strategy. And the problem is not that recordings are correlational rather than causal. Perturbing a brain area and seeing an effect on behavior would also not reveal how the animal was making its decision when there was no perturbation. By manipulating the brain we can find out what brain areas *can* affect behavior on this task, but not what brain areas *do* affect behavior as the rat is making its right turn.

Instead, what Packard and McGaugh (1996) did was simple and cheap: they turned the maze around, so that if the base of the T was pointing to the south, now it was pointing to the north. They then set the rat at the base of the T, and observed its behavior. If the rat continued to turn right, that would indicate an egocentric

strategy. If the rat turned left, it must be following the peripheral external cues, and turning to the same side of the room. The latter is indeed what happened for most of the rats after 8 days of training, but not after 16 days of training, suggesting that decision-making strategies change with extended training. This elegant behavioral manipulation thus informed us of a neural strategy, suggested representations that are necessary for executing it (the rat must be representing the external cues), and allowed for a variety of followup experiments that would delineate the conditions under which rodents choose to use an allocentric rather than egocentric strategy. Understanding these conditions then led to computational models that specified the types of data that the animal must learn and use in each condition, and the computations that may support transitioning from one strategy to another (Daw, Niv, & Dayan, 2005). Indeed, neuroscientific experiments involving lesioning or inactivating brain areas continued to reveal some computational principles of this decision making process (see the next section), and the transition between different behavioral strategies has been linked to mental health phenomena such as addiction and obsessive-compulsive disorder (Voon et al., 2015). However, the initial findings were due to cheap, but extremely clever, behavioral experiments.

In fact, the most fundamental observation about learning—that it proceeds through error-correction—was based on behavioral findings. In particular, if you train an animal that a neutral stimulus, say, a flashing light, predicts the availability of food, the animal will learn this relationship, as can be measured from the animal's salivation response or food-cup approach once the light begins flashing. Similarly, if a tone is paired consistently with food, the animal will show the same responses to the tone. However, in a series of carefully-controlled experiments in the 1960s, Leon Kamin showed that if the light is first paired with food until learning asymptotes, and only *then* the tone is added to the light (still with food following presentation of both), the animal will not learn to respond to the tone (Kamin, 1968). This phenomenon of "blocking" showed that presenting two stimuli in a predictable temporal relationship is not sufficient to engender learning. Instead, learning requires a "prediction error"—the motivationally significant outcome (i.e., the food) must not already be fully predicted (in this case, by the light).

Based on these and other seminal behavioral findings, Robert Rescorla and Alan Wagner proposed a computational model of learning based on prediction errors (Rescorla & Wagner, 1972) that is, to this day almost 50 years forward, the most influential account of trial-and-error learning in animals and humans. Together with later computational models of reinforcement learning showing how one can use prediction errors to optimally learn the sum of future rewards predicated on a certain state or stimulus (Barto, Sutton, & Anderson, 1983; Sutton, 1988), and recordings of activity of dopaminergic neurons while monkeys learned to associate cues with motivationally significant outcomes (Ljungberg, Apicella, & Schultz, 1992), these findings led to the influential identification of phasic dopaminergic signals with a reward prediction-error signal (Barto, 1995; Montague, Dayan, Person, & Sejnowski, 1995; Montague, Dayan, & Sejnowski, 1996), widely heralded as the poster-child of computational neuroscience. This success story — a normative computational theory that explains and predicts key neural signals with precision (e.g., Waelti, Dickinson, & Schultz, 2001; Tobler, Dickinson, & Schultz, 2003), drilling in all the way from Marr's (1982) computational level, through an algorithmic solution and to its neural implementation (Niv, 2009; Niv & Langdon, 2016) — began with

behavioral findings (Sutton & Barto, 1990), and is a prime case of behavior and computation informing our understanding of neural mechanisms, not the other way around.

It goes without saying that ideas about stimulus competition and learning through prediction errors, first identified through blocking, have had far-reaching implications for mental health research. For instance, findings showing impaired blocking in schizophrenia (Martins Serra, Jones, Toone, & Gray, 2001) inspired the hypothesis that positive symptoms of schizophrenia may, in part, be due to spurious associations between stimuli that should otherwise have been subject to blocking. Linking this to aberrant dopamine prediction errors provides another explanation for the beneficial effects of neuroleptics (dopamine antagonists) for treating positive symptoms.

More recently, behavioral experiments have demonstrated a distinction between prediction-error signals due to receiving less reward than expected and prediction-error signals due to making a motor error that misses the rewarding target: while the former prediction errors affect learning of the value of the rewarding option, the latter do not, suggesting a neural dissociation between the two types of prediction errors (McDougle et al., 2016). In particular, the findings suggested that not receiving an expected reward due to one's own reach error does not activate the same type of neural prediction errors – a hypothesis supported by followup work using functional neuroimaging of neural prediction-error signals in the striatum (McDougle et al., 2019).

In the domain of memory, the behavioral phenomenon of "retrieval-induced forgetting" shows that when trying to recall an item, memory traces that are similar enough as to compete for recollection but eventually lose the competition are subsequently weakened (Anderson, 2003). For example, imagine learning the word-pairs `fruit-pear`, `fruit-kiwi` and `fruit-apple`. In a later rehearsal phase, you are requested to complete the stem `fruit-pe__` (presumably with the word "pear"). Since "apple" is the quintessential fruit, and moreover, it bears some semantic similarity to a pear (round-ish, palm-sized, tree-growing fruit that is sometimes green), it may come to mind as you recall the word "pear" and compete for that recollection. Behavioral findings show that if the competition is sufficiently strong (but "pear" still wins, as "apple" cannot complete the word stem), the memory trace for "apple" is weakened. As a result, subjects are less likely to later remember the pair `fruit-apple` in a recall test, or to complete the stem `fruit-a____` with "apple"; moreover, they may even tend to not remember the word "apple" in other pairs (Anderson, Bjork, & Bjork, 1994). This effect is item-specific (it does not affect "kiwi," a fruit that was less likely to come to mind and compete with "pear"), it depends on competition for retrieval (just rehearsing the pair `fruit-pear` does not weaken "apple"), and its boundaries in terms of dependence on practice, memory strength, etc., have been extensively characterized (Anderson & Spellman, 1995). The behavioral phenomenon of retrieval-induced forgetting imposes useful constraints on the implementation of memory systems, suggesting specific forms for networks of memories and how and when they are updated. In particular, it has suggested that the function of oscillating inhibition seen in cortical semantic memory networks and hippocampal attractor networks of episodic memory may be to strengthen weak memories and punish competitors (Norman, Newman, Detre, & Polyn, 2006; Norman, Newman, & Detre, 2007) — insight that would be much harder to glean from neural recordings absent the rich behavioral data.

In my own lab, Nina Rouhani recently completed a series of behavioral studies that revealed more about the organizational structure of episodic memories than we could have hoped for even with invasive neuroscientific techniques. Rouhani asked participants to report reward predictions during learning, and subsequently tested their memory for trial-unique visual items (scenes) that had been presented throughout the task. Her results showed that surprising outcomes (either much more or much less reward than expected) increase memory accuracy for items that coincide with the surprising rewards (Rouhani, Norman, & Niv, 2018). The findings were compatible with two different computational and mechanistic processes: surprising events may be encoded more strongly, or they may be better differentiated from other events, making them easier to recall. By introducing sporadic large change-points in reward outcome and testing for recognition priming (the phenomenon of faster recognition memory for an item when it is "primed" by an event that was linked to it in memory storage), Rouhani was further able to show that memory for a highly surprising event is separated from both the preceding and following items, even while serving as a connection node in the overall network of memories (Rouhani, Norman, Niv, & Bornstein, 2020). These results inform our understanding of the neural organization of memories and potential effects of norepinephrine and dopamine on hippocampal encoding, and further suggest that the source of dopamine may not be the same midbrain dopamine that affects learning (Rouhani et al., 2018), all using behavioral paradigms and computational modeling.

Another example is research into the format and structure of working memory (Miller, 1956), in particular visual working memory (Brady, Konkle, & Alvarez, 2011). Here, too, progress on the key question of whether working memory takes the form of a limited number of discrete slots or rather is limited due to a shared resource—questions that seem most reliably about the neuroscientific implementation of a computational component of perception—has relied on behavioral psychophysics research as well as computational modeling (Bays & Husain, 2008; Brady & Alvarez, 2015; van den Berg, Shin, Chou, George, & Ma, 2012; Keshvari, van den Berg, & Ma, 2013; van den Berg, Awh, & Ma, 2014). Characteristically, neuroscientific evidence for hypotheses derived from computational modeling of behavioral findings played a confirmatory role, rather than revealing the phenomenon (Ma, Husain, & Bays, 2014).

Finally, detailed hypotheses about how attention is deployed and controlled have also been inspired by, and tested in behavioral data. In particular, recent work suggests that even as we attend to an object or a location in space, our attention spotlight is not stationary – it moves around briefly, scanning the environment and returning back to the focus, at a frequency of 8 scans per second (Fiebelkorn, Saalmann, & Kastner, 2013). This "blinking of the attentional spotlight" may be related to other 8Hz (theta rhythm) processes characteristic of environmental sampling. It also constrains the search for brain areas that are responsible for deployment of attention. In particular, further neural data from non-human primates implicated the lateral intraparietal cortex and frontal eye fields—two prominent parts of the classic frontoparietal attention network—in both maintaining attention and suppressing shifts away, and periodically shifting attention to rapidly scan the environment (Fiebelkorn, Pinsk, & Kastner, Under Review). Important for our discussion, Fiebelkorn et al.'s (2013) findings that initially revealed the existence of this fascinating cognitive process

did not involve any neural measurements. Moreover, those findings could hardly have been predicted even knowing that other environmental sampling processes operate at 8Hz. Indeed, it seems that no amount of neural data on attention could have led us to speculate this pattern of behavior with confidence, whereas a single behavioral study was sufficient to establish it.

What have we learned from the brain that behavior had not already taught us?

Studying the brain is undoubtedly useful for neuroscience, especially inasmuch as many questions in neuroscience revolve around how different computations are implemented in neural hardware. The question I am posing here is whether neuroscientific measurements are *sufficient* for making progress on understanding cognition, so much so that their import overwhelms that of behavioral findings, and studies that do not employ neuroscientific measurements should be categorized as "only" behavior and considered to not be relevant to understanding the brain. I also ask whether neuroscientific studies are *necessary* for answering some questions in cognition. That is, is it possible that a study can answer a question about the brain absent any neural measurements? I pose these questions not as straw men. Although neuroscience and behavior need not be at two opposing ends of acceptable research—in fact, they are most effective when done in tandem, ideally also with theoretical modeling—unfortunately, in recent years, the rise of neuroscience has been accompanied by a disdain of behavioral research to the point of rejecting its usefulness altogether. Meetings (e.g. *Computational Systems Neuroscience*), journals (e.g., *The Journal of Neuroscience* – a society journal that supposedly represents the whole field) and, perhaps most importantly, funders (e.g., the *National Institute of Mental Health*) all share the implicit or explicit assumption that behavioral studies cannot lead to valid neuroscientific findings, and therefore are irrelevant to understanding the brain. This leads to rejection of abstracts, papers, and grant proposals, sometimes without review, and often with the claim that they are not topically relevant.

However, the list of true discoveries about perception and cognition from neural measurements or perturbations that were not already known from studies of behavior is embarrassingly short. Of course, beginning from the Hodgkin-Huxley model (Hodgkin & Huxley, 1952)—a fundamental breakthrough in understanding how neurons fire action potentials and communicate that is of course not attributable to any behavioral findings—and continuing with many investigations of anatomy, physiology, and recordings of single neurons while animals were performing different tasks, we have learned a lot about the workings of neurons in different areas of the brain. Still, the breakdown of different functions and cognitive processes into modules that can be expected to be realized independently in neural hardware, and that we should therefore be looking for in different areas, is predominantly thanks to behavioral studies of healthy individuals and those suffering from brain damage.

Arguably, the most interesting findings from neural recordings or perturbations are when those are coupled with incisive, hypothesis-driven behavioral experimental designs. It seems that neuroscience alone is not nearly as revealing as are a combination of a telling behavioral design, neural recordings or perturbation of

neural function (e.g., through lesion, inactivation or stimulation), and a computational model that states the hypotheses to be tested precisely. For this combination to reach its peak performance, work on developing and testing behavioral paradigms, as well as computational models, is as important as work on developing new neuroscience methods. My argument is therefore that pure behavioral research is not only critical for understanding the mind – it is also the cornerstone of understanding how mental processes are implemented in the brain.

Assuming that one overarching goal of neuroscience is to understand how the brain processes information, it is relevant to ask: what have we learned about information processing from neuroscientific research alone? Unfortunately, even with the aid of my entire lab at the time I first conceived of this paper (see acknowledgements), I could not come up with more than a very short list of such findings to highlight here. The finding that face perception is special, for instance, was driven by behavioral findings (Kanwisher & Yovel, 2006)—comparing perception of right-side-up and up-side-down faces (R. K. Yin, 1969; Diamond & Carey, 1986)—long before face cells were discovered in infrotemporal cortex (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Of course, determining whether there are specialized modules for processing faces versus other objects is aided by neuroscientific data (Kanwisher, McDermott, & Chun, 1997; Grill-Spector, Knouf, & Kanwisher, 2004; Yovel & Kanwisher, 2004), however, even those distinctions are often easier to make behaviorally rather than neurally (Leder & Carbon, 2006; Robbins & McKone, 2007). Another example, moving from visual perception to higher cognitive functions, is "theory of mind." In this domain, the concept, its breadth, and its development were all successfully studied without looking into the brain (e.g., Wellman, Cross, & Watson, 2001). Neural studies have revealed potential brain areas underlying theory of mind (Saxe & Kanwisher, 2003; Saxe, Moran, Scholz, & Gabrieli, 2006; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), but have not done much to explain the cognitive process of theory of mind as a series of representations and computations over these. Even domain specificity and dissociations between false beliefs and false photos were seen behaviorally through developmental work before their respective neural counterparts were pinpointed (as reviewed in Saxe, Carey, & Kanwisher, 2004).

Obviously, questions such as "what area of the brain is involved in process X" can only be asked at a neural level. However, my claim is that understanding what process X *is*—what computations it embodies—is rarely done at the neural level alone, if at that level at all. In a sense, the term "computational neuroscience" is a misnomer, as computational discoveries about the brain can hardly skip the psychological-behavioral level. Some exceptions can be found in the domain of perception. One can convincingly argue that measuring the receptive fields of neurons at different levels of the hierarchy of visual processing areas in passive, non-behaving animals, has helped explain how visual processing proceeds from building blocks to percepts (although we are still far from understanding even this basic process, and recently computational models, not neuroscientific data, seem to be providing most of the breakthroughs).

But is perception the "highest" cognitive function for which neuroscience without behavior can inform our understanding of cognitive processes? Although few and far between, there are some instances where neuroscientific research has led to insights about higher cognitive functions that would perhaps not be available

otherwise. This has been mostly by way of lesion studies (or, in animals, reversible inactivations) that illuminated the separable components of cognitive processes once thought to be unitary. In memory research, for instance, the groundbreaking case of Henry Molaison (more commonly known as patient H. M.) first suggested a separation between episodic long-term memory and other forms of memory and learning, revolutionizing the study of memory and the hippocampus, and spurring the field of cognitive neuropsychology (Scoville & Milner, 1957; Augustinack et al., 2014). Notably, the findings relied on testing with appropriate behavioral tasks that demonstrated the different dissociations, for instance between forming new episodic memories and learning new skills (Milner, 2005). Similarly, while behavioral studies such as Packard & McGaugh's (1996; discussed above) suggested that animals transition between decision strategies as they become more familiar with a task, lesion studies in rodents paired with theoretically sophisticated behavioral paradigms revealed that both strategies—goal-directed decision making that relies computationally on planning in a mental model of the world, and habitual responses that lean exclusively on past experience (Dickinson, 1985; Dickinson & Balleine, 2002; Daw, 2018; Drummond & Niv, 2020)—are learned in parallel and, in principle, available for use at any given time (Killcross & Coutureau, 2003; H. H. Yin, Knowlton, & Balleine, 2004, 2005; Balleine, 2005)

Other questions regarding apparent (or true) equivalence between categories of events still await resolution by way of neuroscientific measurements. For instance, is not getting an expected reward equivalent to losing money? Behavior suggests that these two situations are similar, but whether they truly are equivalent is a question best answered at the implementational level. Another question of this same flavor is: do we have two antagonistic motivational systems, an appetitive one and an aversive one (Konorski, 1967), or rather is motivation controlled by one system with two poles? This class of questions can possibly only be answered at a neural level, albeit coupled with a suitable behavioral task. Indeed, these are prime examples for which a generic behavioral test (e.g., extinction learning or conditioned place preference) will not be nearly sufficient, and the neural investigations must rely on a clever behavioral task specifically tailored to the question at hand (see below). Finally, questions about social behavior and its reliance on specially-tailored versus general-purpose neural mechanisms are another class of studies that combine all of Marr's (1982) levels. As Lockwood and colleagues convincingly argue, such questions about which of several algorithms is implemented in the brain can be best answered with neural measurements during properly controlled behavioral experiments tailored to the algorithms being contrasted (Lockwood, Apps, & Chang, 2020).

I therefore want to make clear that I am not arguing for the futility of neuroscientific research. I am calling for a true merger between psychological cognitive science (ultimately interested in understanding behavior) and neuroscience, or at least cognitive neuroscience (ultimately interested in explaining the brain). The study of behavior cannot afford to ignore such an important source of information as the brain—why would we be interested in measurements such as response times and eye movements but *not* be interested in accompanying neural measurements? And similarly, the study of the brain must rely on understanding of behavior if it has any chance of making rapid progress.

Clever behavioral experiments allow causal conclusions despite correlative measures

One last point I would like to make, while on this soapbox, regards a related strong bias in neuroscience: the preference for causal manipulations rather than correlational measurements. This predisposition suggests that since an fMRI signal is correlational, it is by design inferior to a technique that manipulates the neural hardware, for instance, optogenetically. This bias similarly renders behavioral measures in humans absent a brain manipulation (for instance, using transcranial magnetic stimulation, or due to a lesion) automatically inferior. However, I believe that clever behavioral designs have two advantages: they allow behavioral studies to sidestep the causation/correlation pitfall and they help us utilize resources judiciously.

We can start by taking a cue from how the brain goes about making sense of the world: indeed, we often infer causal constructs (e.g., in perception: "what I am looking at is a table surrounded by four chairs") from noisy, low-dimensional, correlational measurements (e.g., the two-dimensional images that fall on my retina) together with prior knowledge about the generative process (that is, how we expect different objects to manifest in such measurements). So, inferring causality from convergent correlational measurements is not an ultimate sin.

The brain also utilizes perturbations – we can move our head to get a different image on our retina if an obscured object seems particularly ambiguous. Or we can walk over to the object in question and move it to verify that all the parts that we thought were connected really do belong to the object. Inferences are therefore reliant on correlational methods very commonly, and on costly perturbation methods in extreme cases. We can similarly construe scientific inquiry: in trying to infer the causal structures of the world around us (why does application of a painful stimulus generate long-lasting fear responses? why are humans so prone to assuming, even on first glance, that some people are more intelligent than others?), it would be wise to combine both correlational measurements and causal perturbations.

The next step is to realize that causal manipulations in neuroscience are not limited to silencing or activating a set of neurons, or lesioning a part of the brain. A behavioral task that requires a cognitive process can effectively apply a causal manipulation, turning a neural process (and its underlying hardware) "on" and "off" through changes in task demands. As an example, imagine the N-back task, in which a participant views a series of letter stimuli and has to respond whenever the current letter is identical to the one viewed N trials back (with N being set in advance, for instance, to 2). This task requires constant maintenance of working memory, introducing and removing stimuli from the "recent N items" set, and comparing the current item to the contents of working memory. Changes in task demands could turn working memory mechanisms on or off, depending on the specifics: setting N to 1 would place minimal requirements on updating of working memory as the task of identifying repetitions can be solved even at the level of iconic visual working memory; or the task can be changed to "respond to up-side-down letters," which does not require working memory at all. Of course, task performance requires more than one cognitive process (e.g.,

in the N-back task: reading, memorizing, comparing the current stimulus to the content of memory, deciding on a response), so a clever, well-controlled experimental design is necessary to single out one function and not others. But this is exactly what the rich legacy of experiments in psychology has taught us to do well. The super-power of psychologists is their ability to design behavioral experiments that isolate and manipulate a process noninvasively, within a whole behaving organism.

As illustrated by the examples in prior sections of this paper, this means that using behavior alone, we can investigate even the neural implementation of working memory. For instance, by assessing the capacity of working memory for colors, orientations, or their conjunction, Luck and Vogel (1997) showed that visual working memory was stored at the level of whole objects (that is, in higher order visual perception areas) and not at the level of individual features. More recently, Katus and Eimer (2018) used the quintessential "causal" behavioral manipulation of dual task requirements: they presented participants with visual and tactile stimuli, separately varying the number of items in each modality, and testing for working memory of only one modality in each trial. The two tasks effectively activated the neural mechanisms responsible for visual and tactile working memory, allowing the researchers to assess whether these mechanisms are shared or separate. The behavioral results—no reduction in working memory accuracy for one modality with the increased demands on the other modality—suggest that independent memory storage exists for each modality, and that capacity constraints on working memory do not result from a shared higher-level control process (Cowan, 2010).

Thus, a relatively simple experiment can answer a question about the segregation of different types of information in working memory, without ever measuring neural signals. The equivalent neural perturbation experiment is dauntingly difficult and less incisive: one would presumably perturb the activity of a brain area (e.g., by TMS) and look for effects on working-memory performance. However, finding that the perturbation does not affect task performance would not mean the brain area is not involved in the task (there could be redundancy in the system), and finding that performance is decreased would not specifically imply deterioration of working memory (here, again, one has to design careful control conditions to rule out other cognitive processes that are involved in the task). Finally, it is not clear what perturbation or neural measurement would tell us conclusively whether tactile and visual working memory rely on shared or separate neural substrates. This is therefore another example where a purely behavioral experiment can answer a neural implementation question more readily than neuroscientific measurements and perturbations can.

To be sure, here again I am not suggesting that neuroscientific measurements are irrelevant, or that there is no conceivable neuroscientific experiment that would answer the above question about working memory. What I am arguing is that the opposite is not true: it is not the case that pure behavioral experiments, using clever experimental designs, and behavioral output such as choices and reaction times, cannot possibly answer a question about neural mechanism. To the contrary, the latter may be better suited to answering that question with little time and effort. Considering behavioral measurements as inferior *a priori* is detrimental to neuroscientific progress.

So why do we chronically devalue behavioral work?

I have illustrated above the stark asymmetry between behavioral research and neuroscientific investigation in understanding cognition: behavioral work contributes more to our understanding of the brain than neuroscience has contributed to understanding the mind. Interestingly, this is similar to the immense unidirectional contributions of the fields of machine learning and computational modeling to neuroscience, all while the former fields have learned relatively little from the brain. Perplexingly, in both cases funding decisions—which are essentially prioritization exercises—suggest the exact opposite[3].

This reversed perception of what field contributes meaningfully to another field may be fueled by several misconceptions. One is a prevalent illusion that neurscientific data are in some sense "objective," whereas using a (computational) theory to interpret (behavioral) data is more "subjective" and less scientific. Another, not unrelated widespread misconception is that behavior is "solved" or not interesting, whereas the "real" questions are ones in neuroscience. But we are still very far from explaining behavior and closing shop in all departments of psychological science (Rescorla, 1988). A question remains, then, what is the best way to make progress in understanding behavior? Although behavior is generated by the brain, I have argued that it is not the case that understanding the brain is the best way to understand behavior. Behavior and neural data are informative about different things: if you are interested in understanding behavior (as are many of us in psychology, cognitive science, and neuroscience), you should study behavior.

This observation suggests that we need a prioritization of questions, not of techniques. Once the questions of interest are clearly defined, for instance "what is a reliable diagnostic phenotype of bipolar disorder" (to take a translational example that seems most readily yielding to neural biomarkers), we can consider what are the best techniques to answer the question expeditiously.

Which brings me to the final illusion: that because behavioral work is sometimes easier (and very often cheaper), its findings are worth less than findings from neuroscientific research. This is a common illusion in economics—people (and animals) value an expensive good more than they do a cheap one (even when the goods are identical) and prefer a reward that they have earned through more effort to one achieved more easily (Aronson & Mills, 1959; Clement, Feltus, Kaiser, & Zentall, 2000; Plassmann, O'Doherty, Shiv, & Rangel, 2008). But this preference is irrational. Sunk costs do not actually make something more valuable, only more... expensive. Do we not, therefore, have a moral obligation to do the behavioral experiments—those that may lead to faster answers at a lower cost—first? Concretely, if a series of behavioral tasks, coupled with computational modeling of the cognitive processes underlying behavior on these tasks, can quantify the aspects of bipolar disorder that fall outside the norm defined by healthy patients, should we

[3] As examples, a program officer at the National Institute for Mental Health (NIMH) requested that a colleague withdraw her funding application from consideration prior to review, saying that regardless of reviewers' evaluation, the clearly mental-health relevant research will not be funded as it does not include a neural component; I have similarly been told, on consultation with several program officers, that a computational psychiatry center that focuses on behavioral measures is not of interest to NIMH unless we include neuroscience methods such as fMRI or MEG, despite research so far showing little return for such techniques in understanding mental illness (Roiser, 2015), not to mention in developing clinically-feasible tools for diagnosis and treatment selection.

not pursue this route before subjecting patients to expensive neuroimaging scans, not to mention invasive techniques?

Conclusion

In basic cognitive psychology research, the role of neuroscience has been suggested to be that of "constraining theories of representation and computation" (Cushman, 2020). It is wholly unclear, however, whether the most effective constraints come from neuroscientific data or from behavioral data. The brain is complex and forgiving – it represents many quantities that may be ancillary to a specific function, and can solve a specific problem through several, often redundant mechanisms. As a result, neural data are often only weakly constraining, if at all. It may be our limitation as researchers (working with the cognitive abilities that our neurons afford us), but historically we have distilled more insight into cognitive processes from contrasting behavior in well-crafted experiments than we have from measuring or perturbing the brain. If we are to accelerate the progress of understanding the human mind, we therefore should restore behavior to its rightful place as the firm base of neuroscience (Griffiths, 2015) on which other findings and types of data can rely, but without which one cannot build a thesis.

Acknowledgements

I am grateful to Daniel Bennett, Mingbo Cai, Nicole Drummond, Sarah Dubrow, Valkyrie Felso, Andra Geana, Gecia Hermsdorff, Angela Langdon, Angela Radulescu, Nina Rouhani, Nicolas Schuck, Melissa Sharpe, Yeon Soon Shin, Olga Lositsky, Diksha Gupta, and Tyler Bonnen for helpful discussions in a lab meeting long ago that contributed critical examples and ideas to this paper, to Daniel Bennett, Nathaniel Daw, Tom Griffiths, Nina Rouhani, Geoff Schoenbaum, Yavin Shaham, Melissa Sharpe and Fiery Cushman for encouraging and insightful comments on an early draft.

References

Ahissar, M., & Hochstein, S. (1993, June). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 5718–5722.

Ahissar, M., & Hochstein, S. (1997, May). Task difficulty and the specificity of perceptual learning. *Nature*, *387*, 401–406. doi: 10.1038/387401a0

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of memory and language*, *49*(4), 415–445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994, September). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of experimental psychology. Learning, memory, and cognition*, *20*, 1063–1087.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychological review*, *102*, 68–100.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psycholog*, *59*, 177-181.

Augustinack, J. C., van der Kouwe, A. J. W., Salat, D. H., Benner, T., Stevens, A. A., Annese, J., ... Corkin, S. (2014, November). H.m.'s contributions to neuroscience: a review and autopsy studies. *Hippocampus*, *24*, 1267–1286. doi: 10.1002/hipo.22354

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol Behav*, *86*(5), 717–730. Retrieved from `http://dx.doi.org/10.1016/j.physbeh.2005.08.061` doi: 10.1016/j.physbeh.2005.08.061

Bargmann, C. I., & Marder, E. (2013, June). From the connectome to brain function. *Nature methods*, *10*, 483–490.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232). Cambridge: MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, *13*, 834-846.

Baylor, D. A., Nunn, B. J., & Schnapf, J. L. (1987, September). Spectral sensitivity of cones of the monkey macaca fascicularis. *The Journal of physiology*, *390*, 145–160.

Bays, P. M., & Husain, M. (2008, August). Dynamic shifts of limited working memory resources in human vision. *Science (New York, N.Y.)*, *321*, 851–854. doi: 10.1126/science.1158023

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of vision*, *15*, 6. doi: 10.1167/15.15.6

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011, May). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, *11*, 4. doi: 10.1167/11.5.4

Clement, T. S., Feltus, J. R., Kaiser, D. H., & Zentall, T. R. (2000, March). "work ethic" in pigeons: reward value is directly related to the effort or time required to obtain the reward. *Psychonomic bulletin & review*, *7*, 100–106.

Coltheart, M. (2006, April). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323–331.

Cowan, N. (2010, February). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, *19*, 51–57. doi: 10.1177/0963721409359277

Cushman, F. (2020). Is cognitive neuroscience an oxymoron? In A. J. Lerner, S. Cullen, & S.-J. Leslie (Eds.), *Current controversies in philosophy of cognitive science* (pp. 121–133). Routledge.

Daw, N. D. (2018). Are we of two minds? *Nature neuroscience*, *21*(11), 1497–1499.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral

striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. Retrieved from http://dx.doi.org/10.1038/nn1560 doi: 10.1038/nn1560

Diamond, R., & Carey, S. (1986, June). Why faces are and are not special: an effect of expertise. *Journal of experimental psychology. General*, *115*, 107–117.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *308*(1135), 67-78.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Learning, motivation and emotion* (Vol. 3, p. 497-533). New York: John Wiley & Sons.

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, *30*(15), R860–R865.

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (Under Review). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention.

Fiebelkorn, I. C., Saalmann, Y. B., & Kastner, S. (2013, December). Rhythmic sampling within and between objects despite sustained attention at a cued location. *Current biology*, *23*, 2553–2558. doi: 10.1016/j.cub.2013.10.063

Griffiths, T. L. (2015, February). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. doi: 10.1016/j.cognition.2014.11.026

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004, May). The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, *7*, 555–562. doi: 10.1038/nn1224

Hochstein, S., & Ahissar, M. (2002, December). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791–804.

Hodgkin, A. L., & Huxley, A. F. (1952, August). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, *117*, 500–544.

Kamin, L. J. (1968). "attention-like" processes in classical conditioning. In *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–31).

Kanwisher, N., McDermott, J., & Chun, M. M. (1997, June). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *17*, 4302–4311.

Kanwisher, N., & Yovel, G. (2006, December). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *361*, 2109–2128. doi: 10.1098/rstb.2006.1934

Katus, T., & Eimer, M. (2018). Independent attention mechanisms control the activation of tactile and visual working memory representations. *Journal of cognitive neuroscience*(Early Access), 1–12.

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS computational biology*, *9*, e1002927. doi: 10.1371/journal.pcbi.1002927

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal conrtex

of rats. *Cerebral Cortex*, *13*, 400-408.

Konorski, J. (1967). *Integrative activity of the brain: An interdisciplinary approach*. Chicago: University of Chicago Press.

Leder, H., & Carbon, C.-C. (2006, February). Face-specific configural processing of relational information. *British journal of psychology*, *97*, 19–29. doi: 10.1348/000712605X54794

Lerner, A. J., Cullen, S., & Leslie, S.-J. (2020). *Current controversies in philosophy of cognitive science*. Routledge.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol*, *67*(1), 145–163.

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020, October). Is there a 'social' brain? implementations and algorithms. *Trends in cognitive sciences*, *24*, 802–813. doi: 10.1016/j.tics.2020.06.011

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279.

Ma, W. J., Husain, M., & Bays, P. M. (2014, March). Changing concepts of working memory. *Nature neuroscience*, *17*, 347–356. doi: 10.1038/nn.3655

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Martins Serra, A., Jones, S. H., Toone, B., & Gray, J. A. (2001, March). Impaired associative learning in chronic schizophrenics and their first-degree relatives: a study of latent inhibition and the kamin blocking effect. *Schizophrenia research*, *48*, 273–289. doi: 10.1016/s0920-9964(00)00141-9

McDougle, S. D., Boggess, M. J., Crossley, M. J., Parvin, D., Ivry, R. B., & Taylor, J. A. (2016, June). Credit assignment in movement-dependent reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 6797–6802. doi: 10.1073/pnas.1523669113

McDougle, S. D., Butcher, P. A., Parvin, D. E., Mushtaq, F., Niv, Y., Ivry, R. B., & Taylor, J. A. (2019). Neural signatures of prediction errors in a decision-making task are modulated by action execution failures. *Current Biology*, *29*(10), 1606–1613.

Miller, G. A. (1956, March). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*, 81–97.

Milner, B. (2005, September). The medial temporal-lobe amnesic syndrome. *The Psychiatric clinics of North America*, *28*, 599–611, 609. doi: 10.1016/j.psc.2005.06.002

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, *377*, 725-728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*(5), 1936-1947.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of neuroscience*, *35*, 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-

sensitive reinforcement-learning process in the human brain. *J Neurosci*, *32*(2), 551–562. doi: 10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67 - 73. doi: http://dx.doi.org/10.1016/j.cobeha.2016.04.005

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006, Jul). How inhibitory oscillations can train neural networks and punish competitors. *Neural computation*, *18*(7), 1577–1610. Retrieved from `http://dx.doi.org/10.1162/neco.2006.18.7.1577` doi: 10.1162/neco.2006.18.7.1577

Norman, K. A., Newman, E. L., & Detre, G. (2007, October). A neural network model of retrieval-induced forgetting. *Psychological review*, *114*, 887–953. doi: 10.1037/0033-295X.114.4.887

Packard, M. G., & McGaugh, J. L. (1996, January). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of learning and memory*, *65*, 65–72. doi: 10.1006/nlme.1996.0007

Plassmann, H., O'Doherty, J., Shiv, B., & Rangel, A. (2008, January). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 1050–1054. doi: 10.1073/pnas.0706929105

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107. doi: 10.1038/nature03687

Rescorla, R. A. (1988, March). Pavlovian conditioning. it's not what you think it is. *The American psychologist*, *43*, 151–160.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robbins, R., & McKone, E. (2007, April). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*, 34–79. doi: 10.1016/j.cognition.2006.02.008

Roiser, J. (2015). What has neuroscience ever done for us? *The Psychologist*, *28*(4), 284–287.

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1430.

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual review of psychology*, *55*, 87–124. doi: 10.1146/annurev.psych.55.090902.142044

Saxe, R., & Kanwisher, N. (2003, August). People thinking about thinking people. the role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*, 1835–1842.

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006, December). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social cognitive and affective neuroscience*, *1*, 229–234. doi: 10.1093/scan/nsl034

Scoville, W. B., & Milner, B. (1957, February). Loss of recent memory after bilateral hippocampal lesions.

Journal of neurology, neurosurgery, and psychiatry, *20*, 11–21.

Stiles, W. S. (1959). Color vision: the approach through increment-threshold sensitivity. *Proceedings of the National Academy of Sciences*, *45*(1), 100–114.

Stiles, W. S., & Burch, J. M. (1959). Npl colour-matching investigation: final report. *Optica Acta: International Journal of Optics*, *6*(1), 1–26.

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning*, *3*, 9-44.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (p. 497-537). MIT Press.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, *23*(32), 10402-10410.

van den Berg, R., Awh, E., & Ma, W. J. (2014, January). Factorial comparison of working memory models. *Psychological review*, *121*, 124–149. doi: 10.1037/a0035234

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012, May). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 8780–8785. doi: 10.1073/pnas.1117465109

Voon, V., Derbyshire, K., Rck, C., Irvine, M. A., Worbe, Y., Enander, J., ... Bullmore, E. T. (2015, March). Disorders of compulsivity: a common bias towards learning habits. *Molecular psychiatry*, *20*, 345–352. doi: 10.1038/mp.2014.44

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43-48.

Wang, X.-J. (2008, October). Decision making in recurrent neuronal circuits. *Neuron*, *60*, 215–234. doi: 10.1016/j.neuron.2008.09.034

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, *72*, 655–684.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of the dosolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 505–512. Retrieved from `http://dx.doi.org/10.1111/j.1460-9568.2005.04219.x` doi: 10.1111/j.1460-9568.2005.04219.x

Yin, R. K. (1969). Looking at upside-down faces. *Journal of experimental psychology*, *81*(1), 141.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010, April). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 6753–6758. doi: 10.1073/pnas.0914826107

Yovel, G., & Kanwisher, N. (2004, December). Face perception: domain specific, not process specific. *Neuron*, *44*, 889–898. doi: 10.1016/j.neuron.2004.11.018