UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PHILOSOPHICAL DIMENSIONS OF THE MORRIS WATER MAZE

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

PHILOSOPHY

by

Jordan B. Dopkins

December 2023

The Dissertation of Jordan Dopkins is
Approved:

_____
Professor Nico Orlandi, Chair

_____
Professor John Ellis, Chair

_____
Distinguished Professor Paul Roth, Ph.D

_____
Assistant Professor Jason Samaha, Ph.D

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

Table of Figures/Illustrations

Abstract:

Philosophical Dimensions of the Morris Water Maze

Jordan Dopkins

In 2014, John O'Keefe was awarded the Nobel Prize in Physiology or Medicine for his work on the hippocampus and its role in encoding map-like representations. His contributions were significantly influenced by Morris water maze studies. O'Keefe himself acknowledged the pivotal role of the Morris water maze, stating that it remains the preeminent behavioral assay for assessing hippocampal function. Indeed, thousands of researchers have turned to the Morris water maze for evidence about navigation abilities and the effects that stress, lesions, pharmaceutical interventions, and more can have on them. This body of studies constitutes an important scientific enterprise.

This dissertation is about some of the philosophical dimensions of Morris water maze studies. Chapter 1 is about the different types of information states (including representational states) found across hypotheses about Morris water maze performance. Chapter 2 is about systematic task failures reported in Morris water maze studies. I argue these impose a constraint on what can count as an explanation (a good hypothesis) of task success. While this has the air of a chopping block for scientific hypotheses, I see things a little differently. In Chapter 3, I argue that satisfying the constraint is a formidable challenge for *any* hypothesis. One that should make researchers second-guess the concepts and

strategies employed to explain the roles information states play in the maze task. Chapter 4 focuses on two spatial concepts. I argue that researchers working definitions for the spatial concepts of *distal* (far) and *proximal* (near) in maze studies are problematic. They are not ecologically valid, and so claims about them do not generalize to real-life navigation behaviors like migration or scavenging behaviors. Following this, I present alternative definitions in terms of neural information about visual cues. The neural information relevant to this account is non-conceptual, and so it provides a sketch of the ways in which information states can fruitfully contribute to explanations of rat success and failure while maintaining ecological validity.

To sum, this dissertation navigates some important philosophical dimensions of Morris water maze studies, illustrating the challenges and opportunities involved in using information states to explain and understanding animal navigation.

Acknowledgments

Thank you, Nico and Jon, for being the best interlocutors and mentors I could have asked for. Your guidance provided me a thoughtful philosophical audience to engage with. Thanks to Paul for sharing your enthusiasm and insights into Sellars. The influence, though subtle, permeates this dissertation. Jason, your expertise in vision science and your philosophical perspective have been instrumental in shaping my understanding.

I am indebted to my comrades in philosophy- Aaron Franklin, Philip Groth, Alea Grundler, Mariana Imaz-Sheinbaum, and Asil M. Martinez. Our conversations have been a crucible for refining my ideas.

Finally, I thank my family for their love and support. Mom and Dad, you never pressured me to do anything else. Thank you so much for that. Amanda, your love, support, and insightful discussions about writing are my constant inspiration. Forest and Hunter, your companionship during countless writing sessions is truly cherished.

0. Introduction

In 2014, John O'Keefe was awarded the Nobel Prize in Physiology or Medicine for his work on the hippocampus and its role in encoding map-like representations. His contributions were significantly influenced by Morris water maze studies. O'Keefe himself acknowledged the pivotal role of the Morris water maze, stating that it remains the preeminent behavioral assay for assessing hippocampal function. Indeed, thousands of researchers have turned to the Morris water maze for evidence about navigation abilities and the effects that stress, lesions, pharmaceutical interventions, and more can have on them. This body of studies constitutes an important scientific enterprise.

While these studies significantly contribute to our understanding of physiology and medicine, they also offer valuable insights into the mind and mental states. Indeed, *The Hippocampus as a Cognitive Map* opens with a 64-page chapter on philosophical and psychological hypotheses about memory, an organism's sense of place, and mental representation (including sections about Kant, Hume, and Berkeley).

Consider that the studies provide a wealth of evidence. Rats are inexpensive, easy to handle and train, relatively easy to affix neural recording equipment to, well-understood genetically and biologically, their central nervous system seems similar enough to those of other mammals, and there seems to be

little moral objection to using them in scientific experiments.[1] As a result of this, there is a lot of neurological and behavioral evidence about rats performances in experiments like those involving Morris Water Mazes; there isn't anything near that amount of data and evidence for other species like humans or primates.[2] What's more is that good hypotheses about the processes and abilities involved in completing the maze task must *fit* this evidence, and so they tend to be rigorous, empirically supported, and cohesive with other explanations in the sciences. As a result of this, they seem like prime candidates for extrapolation to other explanatory projects concerning other processes and performances or even other organisms like primates and humans.[3]

To illustrate, O'Keefe and Nadel's (1978) explanation that rats succeed in later trials of Morris Water Maze experiments because the rat hippocampus functions to construct a cognitive map that the rat uses to navigate its environment. This explanation fits a large amount of evidence collected in experiments, like evidence that rats can navigate mazes in the dark or from different starting locations. O'Keefe and Nadel's explanation fits this evidence since rats navigate using a map, not visual guides. It also fits neurological evidence that neuron cells in the CA1 and CA3 region of the hippocampus display a special kind of activation pattern when the rat is in a specific location in the

[1] For some moral objections see Kitcher, 2015 and Singer, 1977.
[2] For more on this point, see Crystal, 2013 and Chiba, 2015.
[3] For some challenges to this type of extrapolation see Grieves, 2020 and Andrews and Monsó, 2021

maze. O'Keefe and Nadel explain that these activation patterns function like points on a map and represent the rat's present location relative to other locations. Because of this fit, the explanation enjoys wide application to a number of other explanatory projects like human and primate navigation strategies (O'Keefe, 2014), explanations of memory in rats and other species (Redish, 1999; Jeffery, 2018), explanations of hippocampal function, and explanations of sender/receiver roles in the central nervous system (Millikan 2013; Godfrey-Smith, 2013).

This dissertation is about some further, philosophical dimensions of Morris water maze studies. Chapter 1 is about the different types of information states (including representational states) found across hypotheses about Morris water maze performance. Chapter 2 is about systematic task failures reported in Morris water maze studies. I argue these impose a constraint on what can count as an explanation (a good hypothesis) of task success. While this has the air of a chopping block for scientific hypotheses, I see things a little differently. In Chapter 3, I argue that satisfying the constraint is a formidable challenge for *any* hypothesis. One that should make researchers second-guess the concepts and strategies employed to explain the roles information states play in the maze task. Chapter 4 focuses on two spatial concepts. I argue that researchers working definitions for the spatial concepts of *distal* (far) and *proximal* (near) in maze studies are problematic. They are not ecologically valid, and so claims about them do not generalize to real-life navigation behaviors like migration or scavenging behaviors. Following this, I present alternative definitions in terms of neural

information about visual cues. The neural information relevant to this account is non-conceptual, and so it provides a sketch of the ways in which information states can fruitfully contribute to explanations of rat success and failure while maintaining ecological validity.

Chapter 1: Information States and the Morris Water Maze

The Morris water maze is an unspectacular maze. It has no moveable walls, rooms, or chambers**.** It is, basically, a cylindrical tub filled with water. Organisms must find their way out by finding a clear-plastic escape platform hidden just below the water surface. To ensure the platform remains hidden, researchers use solvents to opacify the water.[4] Because of the tub's circular shape, opaque water, hidden reward (safety platform), and water environment, there are no obvious perceptual landmarks available to the laboratory organism (Fig. A). As we will see, that is one reason for the maze's popularity.[5]

Fig. 1. Diagram of a Morris water maze. Depicts a mouse swimming to the clear, center platform. Jones, 2022

The Morris water maze task is relatively straightforward. Researchers take a laboratory organism, usually a rat[6] from its storage pen and place it into the maze. They

[4] The water opacity is crucial to most interpretations of the task, so researchers take it pretty seriously and document their opacifying methods. Early studies used chalk or milk, modern studies use synthetic agents like Lytron 621.

[5] The maze is central to some important research paradigms in cognitive science, neuroscience and biology. For overviews, see Jeffery 2018, Othman, 2022, and Sullivan, 2010. For Morris's original study involving the maze, see Morris 1984. For a detailed guide of this protocol that touches on details specific to actually running a Morris water maze experiment see Vorhees and Williams, 2006.

[6] More is known about mice at the genomic and neurological level, but rats are the preferred laboratory animal for Morris water maze experiments because of they are less anxious and strong swimmers. Whishaw and Tomie (1996) compare mice performances in land mazes to water mazes and report that poor performance is due to poor swimming abilities. Francis et. al. (1995) report poor performances in mice and voles due to anxious behavior when placed in the maze. Studies involve anywhere from 3-100 rats. They are never wild animals. They are obtained from animal model facilities like Taconic Biosciences or Charles River Laboratories. The rats obtained for a study tend to be genetically similar (unless the experiment calls for genetic diversity), and they come from genetic strains like the Sprague-Dawley or Fischer 344 strain,

observe its behavior and/or its neural activity as it tries to find the platform so it can escape the maze. If the rat does not complete the task in 60s,[7] the attempt is counted as a failure. Researchers guide the poorly performing rat to the platform or take it out of the maze to start a new trial. While the number of trials varies from study to study, they typically involve 6-12 trials from random start positions around the maze.

Rats get pretty good at the task. In their first attempt, they take long, inefficient routes that reflect searching behavior more than navigation behavior. The rats cling to the walls of the maze, swim in circles, and finally make their way to the center. By the fourth trial, they do much better. They circle around or change position a bit, then find a straight line to the platform. Finally, by trial eight, they can B-line to the center from any start of the randomized start locations (Fig. 2 and Fig. 3).[8] By every metric, they do really well at the task.[9]

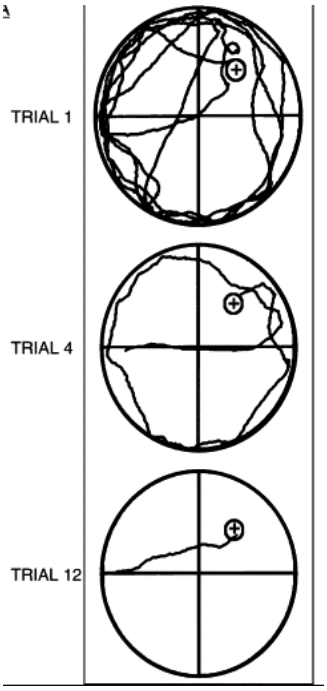
---

which exhibit desirable traits like being easy to handle (Sprague-Dawley) or susceptible to drug addiction (Fischer 344).

[7] In early Morris water maze experiments, researchers would stand near the maze and watch the rat perform the task. In some cases, they would use fairly large and invasive neural recording equipment to monitor the cell activity of individual or clusters of neurons. For detailed accounts, see Morris 1984, O'Keefe and Dostrovosky 1971 & Ranck 1973. Now, researchers try to leave the room. Small cameras are affixed above the maze instead, and less clunky and cumbersome neural recording equipment is affixed to the rat. In addition, proprietary tracking systems like this one by Noldus are used to plot and timestamp rats' locations as they perform the task. The technologies for recording neural activity are less invasive as well. For a good, recent example, see Grieves 2020

[8] I emphasize that the behavior here is not the same as remembering a route from the same start position, like remembering how to get from your door to work. Nor is it the same as remembering and following a list of directions. We are talking about starting from a new location, figuring out where something is, and taking a straight line to it. We are talking about navigating, not remembering directions. Fairly intelligent behavior!

[9] Some important metrics include *Escape latency*-the amount of time (s) the rat spends in the maze. *Swim Path*- the distance (m) the rat swims in a trial. *Thigmotaxia measure*-the amount of time (s) the rat spends circling the outer walls of the maze. For a critical review, see Gehring (2015).



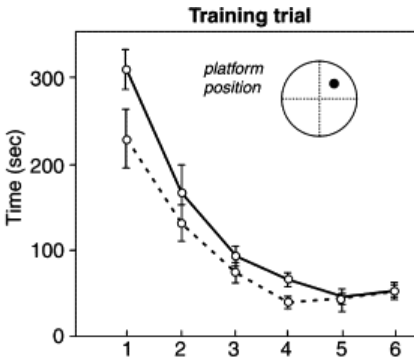


Fig. 2. Swim paths of a rat in a Morris water maze study. Safety platform marked with cross. In Trial 1, the rat swims around the maze a few times before swooping in to find the platform. In Trial 4, it circiles the platform once before finding the platform. In Trial 5, it swims directly to the platform. **Fig. 3.** Average *swim latencies* (time in sec.) for rats in standard maze task. By Trial 6 they substantial decrease their swim latency. Oostra and Nelson 2006.

This is an oft repeated result. In study after study, rats learn to take novel, efficient routes to safety with little, to no perceptible landmarks available for navigation cues. This is regarded as strong evidence the rats have a kind of intelligence. They have abilities to store, process and draw upon information about the maze and platform location. However, the jury is out with respect to the details about their intelligence. Researchers offer a wide variety of hypotheses about the capacities, information, and body parts rats use to succeed at the task. Here are just three examples.

Hypothesis 1: Rats follow the odor trail they left during training. Rats have a pretty good sense of smell, so they just follow their nose to the safety platform (Means, Alexander, and O'Neal, 1992).

Hypothesis 2: Rats see a triggering stimulus like a visual landmark placed outside of the maze, or a stray mark inside the maze wall, associate the view of the landmark with a direction of travel ("if the visual landmark is at 12'0clock in my visual field, then I associate that with the platform being at 3'oclock") and just swim in that direction of travel (Alyan, Touretzky and Tabe, 1995 & Redish, 1999).

Hypothesis 3: Rats build a cognitive map of the maze during training and remember the location of the platform. When placed in the maze, the rat determines its own location from visual cues, determines the quickest route to the platform using its cognitive map and swims in that direction (O'Keefe and Nadel, 1978).

The first hypothesis tells us rats solve the task by using information about things they have direct sensory contact with; they just follow their nose to the platform location. The second tells us rats associate or 'remember' travel information with nearby visual landmarks. They see a triggering stimulus and information about where to go just 'pops into their head'. You can imagine it seeing a visual landmark and automatically thinking something like "if the visual landmark is at 12 O'clock in my visual field, then the platform is located at 3 O'clock. So I must turn and swim in that direction". The third tells us rats construct a cognitive map of their environment during training, which they use to plan and follow routes irrespective of particular perceptible landmarks. Rats

determine their own location in the maze, think about their platform location relative to their starting position and plot a course. No need for a visual landmark to trigger the relevant travel information; they recall or plan it out on their own. Some researchers even think the rats are busy planning out routes while they are tucked away in storage pens.[10]

It is easy to see these are widely divergent hypotheses that lead to very different interpretations of how rats complete the maze task and their underlying thought processes. A few notable differences include the different navigation abilities mentioned across hypotheses[11] or differences between the brain-regions involved.[12] The purpose of this chapter is to focus specifically on the different types of information states mentioned across hypotheses.

This chapter is broken into three sections. I start Section One by explaining what researchers mean by "information" and "information states". Then, I explain two distinctions philosophers of mind use to characterize a variety of things associated with information and representations like thoughts, beliefs, perceptions, images, and sentences. These are distinctions between (1) egocentric and allocentric frames of reference, (2) acquaintance-dependent and acquaintance-independent information states. In Section 2 I draw on these distinctions to characterize some of the different types of information states

[10] Villarreal-Silva et. al (2022) argue that younger rats perform better in the maze because of their memory and planning abilities. These degrade in older rat, and so they perform worse and employ other navigation strategies. Stachenfeld et al (2017) see neural patterns in sleeping rats that resemble those of navigating rats as evidence that rats are dreaming of the maze.
[11] See, for example, Sullivan (2010), Burge (2010), and Rodrigo (2002)
[12] O'Keefe and Nadel (1978) and Reddish (1999)

researchers rely on to explain success in the Morris water maze. This yields a classification of hypotheses. I should note it is not an exhaustive taxonomy. My goal is just to characterize and differentiate some of the information states mentioned in leading (or otherwise important) hypotheses. In section 3 I explain how the proceeding chapters contribute to the landscape of scientific hypotheses characterized throughout this chapter.

Section 1

Researchers use the terms "information" and "information state" loosely for a wide variety of states with a smorgasbord of semantic features. The terms are used for states with intentionality, phenomenology, representational content, tracking functions, indicating functions, or information as described by Stampe (1977) or Dretske (1981).[13] .[14] To see what I mean, it is helpful to look at some examples of how the terms are used and talked about across hypotheses.

In a hypothesis about the role of the ventral striatum, Redish (1999) describes activation patterns of neurons, emotions, and qualitative local view states as all carrying the same information. This use of "information state" is typical and cuts across the meta-semantic differences between these types of states philosophers seem to be interested in.. Someone interested in these differences might say "look,

[13] Stampe (1977) says that a state has information if it is lawfully caused by some other states (Stampe, 1977) and Drestke (1981) says that it has information if it covaries in law-like ways with some other state.
[14] For an account of the meta-semantic differences between these states see Artiga (2016) and Neander (2017). For accounts of how terms like "representation" or "information" are used in brain-sciences see Villaroya (2017) and Rupert (2011). Villaroya conducted a survey polling 100 researchers how they use these terms and published the result. Rupert describes the way these terms are used in scientific explanations of intelligent behavior.

the way the neurons 'carry' that information is surely different than the way a perceptual experience carries that information". It also cuts across epistemic differences between the states. An activation pattern that carries information about x or about the fact that 'x is t' would not fill the same epistemic roles as emotions or sensory states that carry information about x or the fact that x is t. I am not justified in believing "the mug is yellow" just because neurons in my head exhibited some activation pattern. And my fear that the mug will fall will likely play a different epistemic role with respect to my belief that 'I ought to move it' than my belief that 'it is too close to the edge of the table'.

Sometimes, researchers use the word "representation" instead of "information", but they don't appear to mean anything pointed by that use. In most cases, the word serves as a stand-in for "information". Reporting on the results of a survey of 102 brain scientists, Vilarroya (2017) explains that "representation" is used just like "information" to mean a wide variety of relations between 'neural components' and 'environmental components'. In just one part of the report, Vilarroya explains that states that point, designate, cause/effect, signal, transform information, reliably correlated with, and have semantic content all get called "representations". Vilarroya writes "Authors extensively use the notion, putting important explanatory weight on it. However, no agreed benchmark against which to assess specific theoretical and empirical claims exists". So, researchers in these disciplines do not split hairs between information and representations and I will not either (at least, not until chapter three).

The differences that do seem to matter to these hypotheses are captured by distinctions I present below. Those are differences in *what* an information state carries information about and *when* an organism can token or is said to have such an information state. I draw on two sets of conceptual tools to carve out those differences.

The first has to do with the notion of *acquaintance*. I use the term "acquaintance" to mean something like sensory contact. Here are some examples of what I mean. Example 1: A rat is acquainted with a cheese reward when cheese reward particles contact the rat's odor receptors. Example 2: A rat is acquainted with a safety platform if its feet are on the platform or if light bounces off the surface of the platform and contacts the rat's visual receptors. One important aspect of this use is that acquaintance with some feature is feature-dependent. Acquaintance with feature X depends on there being feature X in the environment. A retired laboratory rat's dream, hallucination, or memory of the safety platform does not count as acquaintance with the safety platform.

That is what it is for an organism to be acquainted with some feature of its environment. Now I will explain what it is to say an information state is acquaintance-dependent. I will say that an information state is acquaintance dependent if a necessary condition for it being about some feature is the organism's acquaintance (sensory contact) with that feature. If a rat must be in visual, sensory contact with a mark on the wall for its retinas to carry information about the mark, then those information states are acquaintance dependent. By

contrast, any information states involved in the rat's dream of the mark on the wall are not acquaintance dependent.

The second set of conceptual tools has to do with frames of reference, or the ways in which spatial information is presented. Information states involving *egocentric* frames of reference are about relationships between a target organism (the navigating rat) and features of its environment, like a landmark, edge, or texture it can feel. Some examples of egocentric relationships include the direction or heading of a rat relative to a landmark. For instance, Taube, Muller, and Ranck (1990) describe *head direction cells* in the post subiculum that indicate a rat's heading or direction relative to some landmark in its environment.[15] Other examples include states that are about a rat's distance from a landmark, independent of which way it is facing (Gallistel, 1990 & Redish, 1998) or states that encode vector-like information about distance *and* direction relative to some landmark (Georgopolous et. al. 1983).

They are distinct from *allocentric* frames of reference, which are about relationships between two or more features of the environment independent of the target organism. Examples include the relationship between two visual landmarks or the relationship between a visual landmark and some texture of the maze wall. O'Keefe and Nadel (1978) argue that the hippocampus of navigating rats functions to store information about the distances and direction between visual landmarks.

I will now use these distinctions to carve out and explain a variety of information states relied on across hypotheses about rat success in the Morris water maze task.

[15] For more accounts see Taube, 1995 and McClelland et. al 1995.

Section 2

Acquaintance-dependent egocentric information

One important type of information state is an acquaintance dependent state about egocentric relationships. These states carry information about spatial relationships between the navigating rat and some landmark or feature of the maze. Direction and distance are good examples of such relationships, so long as the direction and distance are defined relative to the rat. Think about the way we say things like "the mailbox is ten feet away from me" or "the starting-line is behind me". That's the way these relations are framed. They characterize spatial relations between the navigating rat and something else.

But! The rat must be in direct sensory contact with that something else for the information state to carry that information. The information states are acquaintance dependent, they can only carry information when the navigating rat is in sensory contact with the feature(s) it is about.

Consider head direction cells, like the ones from Dean, Redgrave, and Westby's 1989 study. Dean et. al explain that the activation patterns of neurons in the superior colliculus carry information about a rat's current direction relative to some visual landmark it is acquainted with.[16] These activation patterns carry information that we might express by saying "the rat's heading is *n* degrees relative to *that* landmark". The states then mediate motor system responses to turn *n* degrees relative to that landmark. Dean et. al. explain,

> "A stimulus that appears (for example) 20 deg. left of the fixation point would correspond to an eye movement to the left of 20 deg. amplitude. Such a movement would bring the stimulus to the centre of the visual field, to be inspected by the foveae of both eyes. This is part of the 'orienting' or 'visual grasp' reflex, particularly useful for systems in which foveal processing is much more detailed than elsewhere in the visual field. In non-primate species, eye movements may be supplemented or replaced by orienting movements of the head, whiskers or ears" Dean, Redgrave, and Westby p.137, 1989.

This ability to turn and approach a landmark that one is acquainted with is often referred to as "beaconing" or "orienting" (Burge, 2010 & Redish, 1999). The idea is that rats see the safety platform from their start position in the maze. The superior colliculus carries information about the rat's current direction, relative to the safety platform, and this mediates motor responses to beacon or turn toward the platform and approach it. Of course, this hypothesis depends on rats abilities to see the platform, which is well hidden in nearly all maze tasks.

One other example comes from the top of the chapter. Means, Alexander, and O'Neal's (1992) argue that rats use the same orienting ability mentioned in Dean et. al., but that they beacon towards a chain of odors left in the rat's previous attempts at the task. Recognizing that rats cannot be visually acquainted with the safety platform in standard Morris water maze tasks, Means et. al. explain that rats explore the maze in their first attempts. After reaching the safety

platform in early attempts, they begin to associate odors left behind in previous attempts with the reward of being safely removed from the maze. The rats become acquainted with an odor, use information about their direction relative to the odor, which then mediates motor responses to reorient themselves and approach the odor. They do this with a chain of odors until they reach the platform.

Heavy or exclusive reliance on this type of information yields a simplified rat psychology. It implies the navigating rats do not need sophisticated mental machinery in order to accomplish the maze task. They could be nothing more than stimulus-response mechanisms that token information in response to a triggering stimulus and forget it when the stimuli is taken away. They are no more complex than sunflowers or insects (Tinbergen 1969 & Morgan 2018). Researchers who stick to this type of information are would be honest in saying "those rats aren't as intelligent as we make them out to be, they only use what's right in front of their face". This simplicity can lend curb appeal to these explanations, especially for causal or syntax friendly explanations that try to explain cognition by appealing to stimulus dependent associations or causal 'brain paths' from input stimuli to output behaviors.

However simple, it is widely accepted now that exclusive reliance on this type of information cannot explaining rat success at the standard Morris water maze task. They do not explain how rats navigate to the safety platform when it is hidden or disguised, which is part of the standard Morris water maze protocol. The best approach to this strategy involves identifying other landmarks or features

that the rat associates with the safety platform like odor trails in Means et. al. 1992. However, this strategy doesn't explain performances in probe trials, where the platform is removed completely or when odors are disguised by chemicals or stirring the water (F. Block, 1999). Despite this, they are important to mention because they were thought, for a while, to explain these performances (see, for instance, Means, Alexander, and O'Neal's, 1992) and they mention abilities and states that are important to other explanations.

Acquaintance-Independent Egocentric Information

These states can carry information about egocentric relationships when a navigating rat is not acquainted with the features it carries information about. Suppose you counted the number of steps it took to get to the coffee shop, then thought about it at home later. The thought would be about a relationship between you and the coffee shop, and it wouldn't depend on your present sensory contact with the coffee shop. In those respects, it is similar to the acquaintance independent egocentric information states mentioned throughout hypotheses of rat success.

Such states are not affected by breaks in the rat's sensory contact with the thing the information is about. Breaks do not disrupt the rat's ability to have information about its relationship to that thing. One way this kind of information can be used for success at the task is by updating it with information about the rat's bodily processes (Rodrigo, 2002 & Burge, 2010). If a rat has the information

that it is 30cm from the platform because it sees the platform, it has the information that each step is 2cm, and that it took 2 steps, then the rat can update its original piece of information so that it is 26cm from the platform. It can do this regardless of whether it maintains sensory contact with the platform or not because the update information is only dependent on information about the rat's bodily processes.[17]

Alyan, Touretzky, and Taube (1995) use these information states to explain performances in Morris water tasks, where rats are not taken far from the maze between attempts. If the rats are kept close enough to the maze, Alvan et. al. explain that they can continue to update the information about their egocentric relationship to the safety platform. Whishaw and Jarrard (1996) use this ability to explain why rats swim to the area near the platform in probe trials. They explain that keeping the rats in close proximity to the maze between trials allowed the rats to update their information about their egocentric relationships to the platform with information about their bodily processes. They then use this to successfully navigate to the area near the platform.

The challenges for relying too heavily on this type of information come in explaining successful performances under the standard protocol, where rats are handled by researchers, stored away from the maze area, and where the safety platform is hidden. It is difficult to understand how rats would update their

[17] It may also have this information as a result of sensory contact, in which case it would have two sources for (at least) two information states about the same information. For discussion of the importance of redundant information states see Rupert (2011).

information with information that is not normal to them, like information about the functions of a rat's vestibular system while it is being held and carried by a researcher.

Acquaintance-Dependent Allocentric Information

Information states that about allocentric relationships are about spatial relationships between features of a rat's environment. Examples include information about the distance between a crease on the maze wall and the safety platform or the vector between a crease on the maze wall, the safety platform, and an extramaze feature like a clock on a laboratory wall. The information states are acquaintance dependent and so they depend on a navigating rat's sensory contact with at least one of the features the information state is about.

These states are sort of like the ones a bowler might use to explain to a friend the relationship between two bowling pins. Suppose the bowler says, "*this* pin is about 6 feet from *that* pin". The bowler's statement would be about a relationship between two features of the bowler's environment that are independent of the bowler. And, because it would involve devices of direct reference, the statement would only carry information about the relationship if the bowler or the bowler's friend were acquainted with the pins.[18]

In a rat's natural environment, the information states could be about the

[18] There are some accounts of direct reference that do not require acquaintance dependence (Russell, 1910).

distance between something like a tree and rock. In a laboratory environment or experiment setting, they might be about features of relevant to the task or features of the lab like the distance between the food feeder and the water feeder. The rat learns to associate the information about the spatial relationship between those features with one or more of the features involved in the relation. So, it associates the water or food feeder with information like "the water feeder is three steps from the food feeder". The information states stored in the rat's memory do not carry this information unless the rat is acquainted with either the water feeder or the food feeder. If it's looking at the water feeder (or the food feeder), these information states can carry the information that it is three steps from the food feeder. If the rat is not in sensory contact with the water feeder (or the food feeder) then they cannot because they lack a cause or input.

Hypotheses that rely on this type of information include include Jeffery's (2018 & 2021) explanation that rats swim to the center platform by using information about the platform's spatial relationship to a feature of the environment the rat is acquainted with from the start position. Jeffery explains that rats associate the spatial information about the safety platform relative to the feature of the maze with the information state it gets from its sensory contact with the platform. To illustrate, it might associate information about the distance relationship between an extramaze cue like a coatrack with its view of the coatrack from the start position. Using this information in combination with information about the rat's egocentric distance relationship to the coatrack, the rat

can determine its distance relationship to the platform and determine a route to the platform.

Schallert et. al. (1996) & Day et. al. (1999) offer similar explanations. However, they argue that a rat will associate information about the spatial relationship of the platform with the rat's *local view* from various places in the maze. Local views are the sum sensory information available to an organism from its present perspective (Redish 1999). They explain that rats learn to associate each local view with information about the bodily movements it would have to take from the position affording the view to reach the platform.

The strategy is like the one given in instructions like this: "Do you see that tall building in front of you? Well, I know you can't see it from here, but the coffee shop is 100ft behind it". The statements in the instructions are about allocentric spatial relationships between features of the environment, and they depend on one's sensory contact with the tall building. The information is employed to 'update' your own information about your egocentric relationship to the coffee shop. Before receiving the instructions, you had no idea how far the coffee shop was. Now, you understand that it is, roughly, 100ft behind the building straight ahead. By combing that with information about your distance relationship to the building (suppose you see you are about 500ft away), you can determine that you are about 600ft away from the coffee shop. So, you use allocentric information to update your egocentric information about your relationship to the coffee shop.

While this information is thought to play an important role in rat success, hypotheses that focus too much on this information face a challenge with respect to the associations the states depend on. Rats take much shorter, more efficient paths to the safety platform in their second attempt at the maze, and they just get better from there on out. That implies the rats either learn this association in their first attempt at the maze or they don't rely on the associations at all. This goes against evidence that these associations often involve repeated exposure and conditioning to develop. Quite often these association-heavy hypotheses cite experiments involving special training protocols, like allowing the rat to roam free in the maze before the trial or using a very large safety platform (more than half the maze size) then reducing the size of the platform over several consecutive attempts until it is about normal sized (Schallert et. al. 1996). Despite this, these explanations are well supported by connectionist and computational models.

Acquaintance-Independent Allocentric Information

Another kind of allocentric information state are those that are acquaintance-independent. They do not depend on the rat's sensory contact with at least one of the features the state is about. The state can be said to carry the information even when the rat is looking elsewhere, has its eyes closed, or is dreaming about the maze.[19] The states are untethered to the rat in the sense that they do not refer to the rat, and they are untethered to the features or locations of a

[19]Stachenfeld et al, 2017

rat's environment in that the rat need not be acquainted with those features or locations to have that information.

Some hypotheses that rely on this type of information state include O'Keefe and Nadel's (1978) explanation that the function of the hippocampus is to produce and transform information states so that they form a cognitive map of the organism's environment. They explain that place cell neurons in the hippocampus exhibit selective firing patterns when the rat is in a specific part of the maze, meaning they only fire when the rat is in a given location in the maze.[20] The place the rat is at when the neuron fires is referred to as a 'place field', and they often look like color swatches or smears on a map. Here is a figure (Fig. 4) from O'Keefe (1976) illustrating the place field of a place cell unit O'Keefe names "Place Cell Unit 213-4-2". Place Cell Unit 213-4-2 only exhibited interesting activation patterns when the rat was in the area of the maze greyed out below in Arm B. The greyed out area indicates the 'place field', or the area that the rat was in when the place cell neuron was recorded to have complex spike activations. This would be like a light that turned on whenever you entered the southeast quadrant of your office. The light *only* turns on when you are in the southeast quadrant just like Place Cell Unit 213-4-2 *only* exhibits complex spike activation patterns when the rat is in that location.

[20] Place cell neurons in the hippocampus are said to fire when they exhibit a complex spike activation pattern, which is a significant deviation from the cells baseline activation pattern.

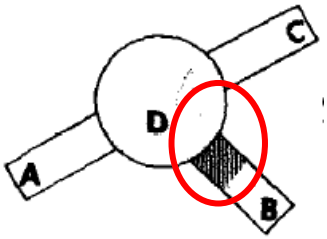


Fig.4. Place field of a single place cell neuron in a T-maze. O'Keefe, 1976

Because of the selective firing, Okeefe and Nadel say that the place field carries information about that place field in the same way that a shaded region on a map carries information about a region in space. Grid cells carry information about relationships between those place fields. We might express that information by saying that the place field J is three feet away or 90 degrees to the left of place field H. They can also carry information about relationships between place fields and features of the rats environment. Many of these information states are about allocentric relationships, like the distance or direction relationships between place fields or features that have nothing to do with the rat. They are also not acquaintance dependent. The information states carry that information even when the rat does not have sensory contact with any one of the things that information state is about. This aspect underwrites rats' abilities to *plan* routes while it is stored away from the maze task or just prior to beginning the task (Stalkonovich 2017). Okeefe and Nadel offer a helpful summary of their view,

> "[W]e think that the concept of absolute space is primary and that its elaboration does not depend upon prior notions of relative space. In our

view, organisms represent space in several independent, though interrelated, ways. A number of neural mechanisms generate psychological spaces referred to the observer, and these are consistent with a relative theory of space. Amongst these are spaces centred on the eye, the head, and the body, all of which can be subsumed under the heading of egocentric space. In addition, there exists at least one neural system which provides the basis for an integrated model of the environment. This system underlies the notion of absolute, unitary space, which is a non-centred stationary framework through which the organism and its egocentric spaces move. We shall call the system which generates this absolute space a cognitive map and will identify it with the hippocampus … A cognitive map would consist of two major systems. The first is a memory system which contains information about places in the organism's environment, their spatial relations, and the existence of specific objects in specific places. The second, misplace system signals changes in a particular place, involving either the presence of a new object or the absence of an old one. The place system permits an animal to locate itself in a familiar environment without reference to any specific sensory input, to go from one place to another independent of particular inputs (cues) or outputs (responses), and to link together conceptually parts of an environment which have never been experienced at the same time. The misplace system is primarily responsible for exploration, a species-typical

> behaviour which functions to build maps of new environments and to incorporate new information into existing maps." O'Keefe and Nadel, 1978. P.1-3.

Applied to success in Morris water maze tasks, the explanation states that rats use the first attempt to gather information about the maze. In later attempts, it goes into the task with a complex system of information states stored during earlier attempts. These function like a map of the water maze. Similar to the way a person uses a map to determine a route from their current location to a destination, so does the rat use this information to determine its position relative to the safety platform and determine a route to the platform.

Redish (1999) offers a similar hypothesis but implicates the hippocampus in the role of long-term memory. The idea is that rats can remember these routes and recall them well after leaving the maze. Redish's explanation helps explain how it is that rats can switch between maps for different environments. Features of the environment call up different maps from memory based on those features. Other examples are found in in Garthe and Kempermann (2013).

While there is some disagreement over the details of O'Keefe and Nadel's explanation, like, for instance, the role that memory plays in navigation strategies and which substructures of the central nervous system implement the information states and transformations involved, the general picture is widely accepted and

applied to a number of other spatial performances in rats and other species like apes and humans.

Information about the Rat's Body

One kind of information that falls out of this distinction-carved landscape is information about the navigating rat's own body, like, for example, information about the number of steps the rat has taken. These states are about processes *internal* to the rat and that are often carried out by organs or subsystems, like the motor systems responsible for initiating a step. They do not carry information about things that are external to the rat like features of its environment or relationships between those features.

Sometimes the states are about processes the rat has already initiated, like the information about steps described above. These can function like a log or store of the rat's past movements (Benhamou, 1997 and Biegler, 2000). Sometimes, they are about processes the rat *will* initiate and are treated like commands or rules like "initiate reorient movement y, then initiate approach motor sequence x" (Rodrigo, 2002 and Stachenfeld et. al. 2017). In either case, the states retain the core feature of being about the rat's own bodily processes and are not about external features.

The distinctions I've used so far have little traction with these states. The egocentric/allocentric distinction is a distinction between two kinds of relationships that information states can be about: egocentric relationships are

between a rat and a feature, while allocentric relationships are between two features independent of the rat. The information states relevant here are not about relationships really; they are about the rat's own bodily processes, and so the distinction doesn't apply. The acquaintance dependent/not acquaintance dependent distinction may apply here, but its application is fuzzy. It is not clear from explanations whether rats are acquainted with their bodily processes and functions via some introspective proprioceptive sense.[21] Perhaps they are, but the information states mentioned here are distinctive enough that they can be characterized without this distinction, and so I will proceed without using it.

One example of a hypothesis that relies heavily on this type of state is Carr and Watson's (1908) hypothesis that rats use a sequence of information states about its own bodily processes to form a route to a reward. Each information state is about a process the rat will perform. So, from it's start position the rat performs the bodily processes from a list of information state. Carr and Watson emphasize that these routes can be very complex. Honzick (1936) used a similar hypothesis to explain rats' abilities to complete a 14-junction maze with speed and 'confidence'. Honzick thought this also explained why rats would crash into walls when the walls were moved or rotated.[22] While the rats use sensory information to guide them, they rely on patterns of stimuli along withinformation about internal

[21] For philosophical accounts of such a sense see Noe, 2004 and Schusterman, 2008. For some examples of explanations involving acquaintance with information states via some introspective sense in like this in ants see Wehner and Flatt, 1972.
[22] Honzick makes sure to emphasize the 'crash'. The rats are running with purpose and confidence in their list of instructions.

processes, and it leads them down the wrong path. Redish (1999) gives a succinct summary of how these types of explanations are used in explaining successful performances in Morris water maze tasks:

> "[I]f the animal always starts facing the east wall at the easternmost point of the environment and the [safety] platform is always along the northernmost part of the wall, then the animal only needs to learn to turn left and then proceed. … [These] strategies generally consist of stereotyped movements independent of starting position." Redish, 1999. P. 11-12.

You can imagine turn-by-turn directions that only mention the driver's movements. Or, you might think about directions given to submarine pilot: **STEP 1**: move the red lever up three times. **STEP 2**: move the green lever to the left twice. They would make no mention of the submarine's bearing or location to anything external to the submarine because the pilot would not have access to that information. They would only mention processes the pilot should initiate.

Hypotheses that rely heavily on these information states do a good job of explaining how rats succeed in the absence of information about its environment, like when it navigates a maze task in the dark (Quirk, Muller, and Kubie 1990). However, they do not explain some frequently replicated behavioral data and experimental results. They do not explain how rats swim from a random (and

likely novel) start location to the safety platform. The starting point is supposed to trigger a sequence of moves that the rat follows to the safety platform. But if the rat starts from a different starting point, the sequence of moves will lead the rat away from the safety platform. Randomizing start locations is a basic feature of the Morris water maze experiment protocol, and rats perform successfully despite the randomization. So, these explanations fail to explain frequently repeated performances that follow the basic experiment protocol.[23]The states are important to mention though because they are often combined with other information states in hypotheses.

For example, when combo-ed with some of the other types of information states, states about the rat's own body are said to underwrite abilities for *path integration*, which is the ability to *return* to a reference point by updating information about egocentric relationships with information about bodily processes. Path integration behaviors involve a familiarity with the destination. Hence the emphasis on "returning" above. The core of this ability is that an organism can continue to navigate to its destination once its lost sight of it (or other sensory contact with it) by relying on information about bodily processes. Paradigm examples of behaviors involving path integration outside of the laboratory include long-distance migration patterns observed in birds (von Saint Paul 1982) and mammals (Darwin, 1873 & Mittelstaedt, 1980). They also include homing behaviors, where animals return home after being displaced, as exhibited

[23] See, for instance Morris, 1981 and Knierim et.al., 2011

by ants (Rodrigo, 2002) and dogs (Séguinot, Cattet, and Benhamou, 1998). These behaviors involve orientation toward a stable or fixed landmark like the sun, stars, or polarized light) and updates via constant processing of bodily information (Rodrigo 2002).

Section 3

There are a wide variety of information states at play across hypotheses about rat success in the standard Morris water maze task. Consider (again) the hypotheses from the top of the chapter.

> **Hypothesis 1**: Rats follow the odor trail they left during training. Rats have a pretty good sense of smell, so they just follow their nose to the safety platform (Means, Alexander, and O'Neal, 1992).

> **Hypothesis 2**: Rats see a triggering stimulus like a visual landmark placed outside of the maze, or a stray mark inside the maze wall, associate the view of the landmark with a direction of travel ("if the visual landmark is at 12'0clock in my visual field, then I associate that with the platform being at 3'oclock") and just swim in that direction of travel (Alyan, Touretzky and Tabe, 1995 & Redish, 1999).

> **Hypothesis 3:** Rats build a cognitive map of the maze during training and remember the location of the platform. When placed in the maze, the rat

> determines its own location from visual cues, determines the quickest route to the platform using its cognitive map and swims in that direction (O'Keefe and Nadel, 1978).

Means et. al tell us the rats rely heavily on acquaintance-dependent egocentric information about odor trails in order to find the platform. Alyan et. al say that the rats use spatial information about features of the maze, but that they require a triggering stimulus to recall that information, and so the rats rely heavily on acquaintance-dependent allocentric information. O'Keefe and Nadel explain the rats' success by pointing to a 'mental' map chock-full of acquaintance-dependent allocentric information. We have three different hypotheses characterized by three very different types of information states.

The rest of this dissertation is concerned with tackling the landscape of information-laden hypotheses about rat success in Morris water maze tasks. Focus on information states promises to be fruitful. We have already seen that exclusive or heavy-handed use of one type of information state can be to the detriment of a hypothesis. How else can a focus on information states help us navigate these hypotheses?

In Chapter 2, I identify a methodological constraint on what can count as a good explanation of rat success. I argue that any hypothesis worth its salt must also be able to explain (identify difference makers to) task failures widely

reported in studies using modified versions of the maze task. The ones that don't should be regarded as mere hypotheses.

This has the air of a scientific chopping block -one that lends an opportunity to weed out bad hypotheses. But I see things a little differently. In Chapter 3, I argue the constraint amounts to a tough challenge for any hypothesis, and that it should make us second-guess many of the concepts and strategies employed to explain the roles information states play in intelligent behavior. In particular, I argue there are significant problems for relying on dislocations (disturbances in the environment), functions, and conceptual representations to explain those failures. At the end of the chapter, I lay directions for future research: detailing how treating information states as non-conceptual information states can help explain the failures.

Chapter 4 focuses on another problem. This one has to do with some of the spatial concepts researchers employ in their hypotheses. Researchers often claim that organisms use distal (far) visual
cues differently than proximal (near) ones. For example, in a behavioral study, Hébert et. al. (2017) report that removing distal cues completely disrupted rodents' ability to complete navigation tasks, while removing proximal cues had no effect. The dominant working definitions for "distal" and "proximal" define those terms relative to the boundaries of the Morris water maze. Distal cues sit beyond the walls of the maze and proximal sit within. The problem with these working definitions is that they do not allow for claims about distal and proximal

cues to generalize to real-life navigation behaviors like long-distance migration because there can be no experimental apparatuses (like mazes) in those environments. I present alternative definitions in terms of neural information about visual cues. The neural information relevant to this account is non-conceptual, and so it provides a sketch of the ways in which information states can fruitfully contribute to explanations of rat success and failure. I hope to give a fully-fleshed out account of that in future work.

Chapter 2: Systematic Failure, Evidence, and Explanation

"… nothing happens without a reason." Leibniz (1677)

In Chapter 1, I described performances of rats in the standard Morris water maze tasks and characterized some hypotheses about those performances. In this chapter, I motivate a constraint on what should count as an *explanation* (a good, successful hypothesis) of those performances.

This chapter is broken into three sections. In section one, I describe task failures from Morris water maze studies involving a few, slight modification to the original maze task. The researchers conducting these studies place a few visual cues outside of the task to help the rats learn the standard maze task explained in Chapter 1, then they introduce a probe trial in which they rotate the extramaze visual cues. Perhaps unsurprisingly, rats fail at the task in the probe trial. I explain the modified experiment protocol and characterize rat performances. In section two, I explain how these task failures motivate a constraint on what should count as an explanation of successful performances in Morris water maze tasks. I explain the constraint and why failure to satisfy it makes a hypothesis bad. In section three, I consider what appears to be the only way of avoiding the constraint, and I argue that it is a dead end. By the end of the chapter, I hope to have demonstrated that good explanations of successful Morris Water Maze performances must satisfy the constraint described in this chapter.

Section 1

Once a rat has learned to take direct, short paths to the location of the safety platform in a Morris water maze, performances in which it fails to do that are counted as failures. Failures are rare in the standard Morris water maze tasks, but they happen occasionally.[24] They are much more common and systematic in experiments that introduce additional protocols to the maze task like stress conditions (Warner et. al. 2013), lesions the central nervous system like hippocampal lesions(Broadbent et al. 2006), or manipulation of features of the experiment setting (Carman et al. 2002 and Stackman et al. 2012).

The kind of task failures I am interested in are reported in experiments by McGauran et al. (2004) and Wortwein et al. (1995). McGauran et al. and Wortwein et al. designed their experiments so that they could manipulate the extramaze cues (stuff located outside of the maze) viewable from inside the maze. To do this, they surrounded the maze with black curtains to cut off view of the laboratory and attached paper shapes to the inside of the curtain, which were designed to serve as the rats' only extramaze cues. They trained rats to do well in the standard Morris water maze task under this design, then they introduced special cue rotation trials, which vary slightly between the researchers' experiments.

[24] Gehrin et al (2015) report two instances of one-off task failures in a control group completing the Standard morris water maze task.

McGauran et al. removed the safety platform and rotated the paper shapes 180 degrees in their cue rotation trials. They recorded the swim paths of each rat and the total amount of time spent near the location of the safety platform during training. They report that every single rat was "clearly impaired" by these manipulations. Instead of swimming to the actual location the safety platform, the rats swam to the place they thought it would be, relative to the rotation of the extramaze cues. McGauran et al report that rats would swim around the incorrect platform area for a while, searching for the platform, then they would initiate explore behavior until the trial ended and the rats were removed. Fig. 5 is helpful for seeing this.

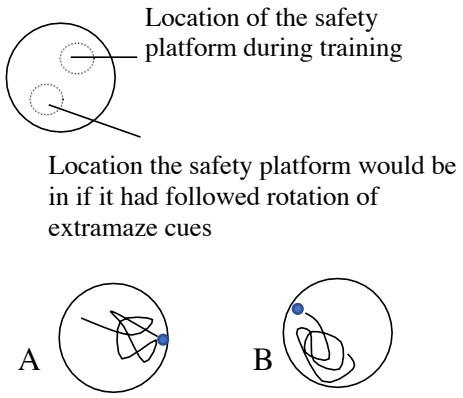


Fig. 5. McGauran, 2004. The top circle shows the actual location of the safety platform during training, and the location the safety platform would have been in had it been rotated with the extramaze cues. A shows the swim path of a rat from a control group, who performed the task without extramaze cue rotation. B shows the swim path of a rat in a probe trial, where the extramaze cues were rotated 180 degrees. Blue dots indicate random start locations.

Fig. 5 shows the swim paths of two rats from McGauran's experiment. The swim path in A came from a rat in a control group that did not have any of their extramaze cues rotated before the probe trial. For these rats, the cues remained constant between training and the probe trial. The rat swam directly

from its start location to the location the safety platform had been during training. Then, it exhibited search behavior in that area before looking for a new way out. Had the platform been left in the maze, it would have escaped quickly and efficiently. Compare this with the swim path in B, which came from a rat attempting the task during a special cue rotation probe trial. The rat swam directly to the location the safety platform would have been had it been rotated similarly to the cues. It exhibited search behavior here, then explored the maze for a new way out. Unlike its counterpart, it would not have escaped quickly and efficiently. The crucial difference here is that the rat in B swam directly to an incorrect location. It went to the place it 'thought' the platform would be in, not the place it actually had been in during training.

Wortwein et al. give slightly modified versions of this task and found similar results. They rotated their paper shapes 90 degrees in one type of trial and 45 degrees in another, and they did not remove the safety platform. They recorded the total swim duration in seconds and distance in pixels that it took for rats to find the safety platform. Wortwein et al. report that rats were "significantly impaired" in both cue rotation trials. Like in McGauran's experiments, every rat appeared to rely on the rotated cues to find the safety platform and swam in the direction of the cues' rotation. They initiated search behaviors in the place the safety platform would have been, had it been rotated at 45 or 90 degrees too, then they initiated explore behavior until they found the safety platform. They report that, on average, rats in cue rotation trials swam almost three times the duration

and distance they did on their last training attempt.

Performances like these have been repeated in similar experiments by Vorhees and Williams et (2006) and Rivard et. al (2004). They can also be found in experiments that make use of other maze types like the radial arm maze (Suzuki et al. 1980) or Barnes maze (Harrison et al. 2006). This is just to say that the effects reported by McGauran and Wortwein are not one-off effects or anomalies; they contribute important evidence about how rats use visual information that has been replicated across experiment designs.[25]

In Chapter 1, I explained that researchers count swimming directly to the location of the safety platform as successful because that behavior is statistically normal or taken to mark some adaptive or learned ability. The performances reported by McGauran et al. and Wortwein et al. are counted as task- failures because they detract from those performances. The rats' swim paths are deviations from the direct swim paths reported in thousands of Morris water maze experiments (and in their control groups), and rats who take them take longer, swim farther, and spend more amounts of time in the wrong parts of the maze.[26] Another way to think about task failures is that were escaping *really* a matter of life and death, the organism would perish or exert significantly more effort to escape. The performances in the cue rotation trials are also importantly different

[25] For a review of these kinds of experiments and the important evidence they provide about navigation abilities and information see Kneirem and Hamilton, 2011
[26] For a detailed account of the quantitative measures used to measure performance and evaluate errors, see Hooge and De Deyn, 2001.

from the rat's first attempts, where it just explores the maze until it finds the platform. In the cue rotation trials, the rat swims directly to a spot in the maze. It's just the wrong spot. This is taken to signal that something was learned or that some non-exploring adaptive ability 'kicked in', but those abilities betray rats in cue rotation trials.

Section 2

In this section, I explain how these task failures motivate a constraint on what counts as good explanations of success performances. But first, a quick note about my use of the word "explanation" is in order. Explanations identify difference makers and help researchers make predictions as a result. Hypotheses, on the other hand, are efforts to determine difference makers and make predictions. "Explanation" is, in other words, a success term and "hypothesis" is not. (Khalifia 2017 and Woodward and Ross 2023).

This signals an important shift in the dissertation. In Chapter 1, I described some hypotheses, and didn't do much to split hairs between good ones and bad ones. I described research that means to identify, among other things, difference makers to rats' successful performances in maze tasks. In this chapter, I motivate a constraint on what should count as a *good,* successful hypothesis about those performances. I motivate a constraint on what should count as *explanations* of those performances. It marks a significant move from describing what's on offer in the sciences already to taking the first steps in the arguments of this

dissertation, which is about what they need to offer in order to count as good science.

The first step involves pointing out that researchers understand the task failures[27] as evidence that rotating the extramaze cues causes something to happen to the information systems relevant to the earlier success performances, and, partly because of this, rats swim in the wrong direction. I emphasize that they understand it to be evidence that *the very same* information system that helps explain success is also meant to make a difference to the task failures. Researchers think that something happens to those systems and the information states they read or write that makes a difference to the performances in those trials. The cue rotation trials are supposed to tell us about how those information systems operate and what sorts of environmental features they are sensitive to.[28] Hence claims from Wortwein et al. and McGauran et al. like the following:

> "We demonstrate that rotations of distal cues and starting position [in probe trials] impair retention of the platform's location. We suggest that

[27] Here, I mean the kind of systematic task failures described by McGauran and Wortwein.
[28] These experiments provide important evidence, conceptual tools, and experiment results for thinking about how different cues are used in navigation and spatial learning. As a result, explanations that draw on research in this paradigm end up being better explanations because they offer more fine-grained accounts of the causes or difference-makers to behavior. For instance, McGauran et al. and Wortwein et al. draw on the performances they report in cue rotation trials as evidence that rats preferentially use information about *distal* (far away, extramaze) cues over information about *proximal (*nearby, intramaze) cues in navigation. Attention to this difference between cues promises to identify more fine-grained sources for rats' navigation or spatial learning performances. Instead of pointing to acquaintance dependent allocentric information states as a source for performance, research could point to acquaintance dependent allocentric information states *about distal cues or proximal cues* as a source for performances.

> the association between the configuration of distal cues and platform location is retained in memory but the association is fragile and sensitive to disruption … Rotation of the cues 180° clearly impaired the search strategy of this group". McGauran et al. 2004

> "[T]he rats' performance on the rotation sessions demonstrated [rats] to be significantly impaired by the 90" rotation of the distal cues. As discussed in the Introduction this result indicates that [rats] discriminated between the individual cues and utilized such a discrimination for navigational purposes [in both standard and rotation trials]." Wortwein, et. al. 1995

McGauran et. al point to impairment in the rat's strategy and a resulting task failure. They think rotating the extramaze cue screws up the rats' plans. And Wortwein et al point out that the same discriminations used in standard and probe trials yield different performances.

From where things stand after this first step, it seems that any explanation of success performances in Morris water maze tasks faces a constraint: **the task-relevant information system(s) from the explanation must also make a difference to the kinds of systematic task failures described by Wortwein and McGauran.**[29] For example, if visual information about a landmark is said to

[29] The success-relevant information systems may also need to make a difference to other kinds of systematic task failures reported in other studies, but I am focusing on their need to make contributions to the specific failures reported by Wortwein and McGauran.

explain success, then visual information about the landmark must also explain the task failures in the probe trials. That is to say that it must be a *difference maker* to the task failures.

One way to be a difference maker is to have some effect or be part of a chain of causes of the failures. I say that because it's a helpful way of thinking about what it means to be a difference maker, and so sometimes I'll put a point that way. At the very least though, the task-relevant system should impact the task failures so that the performance would have been different if the system weren't part of the story- there should be a difference in those counter-factual situations. That is not to say that the success explanation must actually include a second explanation (like a written or typed explanation) of how the information system makes a difference to the task failures. The constraint is much weaker than that. It constrains the information systems identified in an explanation of success so that the system must be a difference maker to the failures, even if no account is given of the differences it makes to the failures. The idea is to leave the door open for such explanations, not to require that they are actually given.

Failure to satisfy this constraint spells trouble for a hypothesis about success. It means it cannot achieve the status of explanation, or, to put the point crudely, it is not a good hypothesis. Here's why, if the information systems from a hypothesis about success do not make a difference to the task failure, then they do not explain the failures (as I mentioned earlier, being a difference maker is necessary to explanation in these paradigms). The cue rotation experiments are

taken to demonstrate, that the information systems that explain success also explain the failures. So, if the information system from an explanation for success does not make a difference to the failures and so does not explain the failures, then it does not really explain success. Again, something is thought to happen to the information systems responsible for success in those probe trials that explains and, as a result, makes a difference to the failures.

To see what I mean, consider Means, Alexander, and O'Neal's (1992) hypothesis that rats detect odor trails left during training, and that it follows those to successfully navigate the maze task. Since there are no odor trails leading to the wrong location the rat swims to in the cue rotation trials, there is nothing for the rat to detect and follow to that location. The rat's ability to follow those trails cannot be a cause or difference maker to the behavior. In fact, it seems that had the rats in cue rotation trials followed their nose, they would have performed successfully! It follows that the information systems responsible for following an odor picked up during training are not difference makers to the task failures and so they do not explain the task failures. And because they do not explain the task failures, they do not explain the success performances. If it seemed like they explained the successes it was likely because we were ignoring evidence about the sorts of performances those mechanisms are expected to produce in cue rotation trials.

Section 3

In Chapter 3 I will consider three strategies for satisfying the constraint and some of explanatory challenges associated with the strategies. However, before I do that, it is worth considering a way of avoiding the constraint altogether. On this way of thinking, hypotheses about success only *seem* to face a constraint in light of the task failures reported by Wortwein et al. and McGauran et al. But! Rather than take those as evidence that the same information systems explain successes and failures, researchers should take them as evidence that there is a distinct system that explains the failures and is separate from the one that explains successes. The success and failure performances are sort of like the PRINT and FAX functions of a Xerox machine in that they get explained by different mechanisms that just happen to be housed in the same beige box. There are a few ways this could go in explaining the rats' task failures. The failure producing system could be said to 'kick in' and replace the success producing system, like the way second gear kicks in to replace first gear in powering a car. Or it could be said to 'win out' in competition with that system, like the way a stronger radio signal wins out over a weaker one to produce the sounds coming out of my radio.[30] Whatever the case, the system that explains the failures is different from the one identified in the explanation for success.

[30] Figuring this out assumes the hypotheses are separate, and so the question of how the hypotheses hang together at the basic level I discuss here is prior to questions about whether one of the different sources wins out in competition or replaces the other. That question about how the hypotheses hang together comes in at a slightly less basic level.

Why would someone pursue this line of thinking? I see two reasons, but I emphasize that, under scrutiny, they don't turn out to be *good* reasons, just reasons. They're like legal precedent that doesn't pan out as good grounds for a legal strategy. Sure, it's legal precedent, but it's not *good* legal precedent.

The first reason comes from recognition that organisms often learn compensatory mechanisms to help them perform tasks in case their primary mechanism for it fails. As a side note, there's an interesting scientific debate about how to think about these compensatory mechanisms that is underexplored by philosophers. At the core of the debate are philosophical questions about how to think about the relationship between brain structures like the hippocampus, whose physiological function is closely associated with a cognitive function, like building a cognitive map of the environment. When those brain structures are damaged in lesion studies, organisms perform worse on tasks at first, but return to close levels of original performance after repeated exposure to the task (Mogensen and Mala 2009). The question is whether compensatory mechanisms should be thought of as involving some other brain structure that steps in to perform the function of the damaged structure or whether the compensatory mechanism involves incorporating another cognitive strategy altogether. To put this crudely, when iMaps fails on my phone, I can launch the same program on another device, or I can resort to a new strategy like using a compass. The compass uses different information and different 'hardware'. Are compensatory mechanisms akin to launching iMaps on another device? Or resorting to a map

and compass? Here is why I think this debate is interesting for philosophers. The evidence from research on compensatory mechanisms in hippocampal map paradigms and the ways researchers think about how best to explain it bears interesting connections to some traditional philosophy of mind topics like localization of cognitive function (Fodor 1983, Mogensen 2011) or multiple realizability (Polger and Shapiro 2016). For example, evidence of degraded performance after learning seems to count as evidence that different strategies are being employed, which wouldn't count as evidence against localization or for multiple realizability of cognitive function. But if, on closer look, or on other tasks, performance isn't degraded, that would seem to count as evidence that the same strategy is being employed by other brain structures (In a separate experiment from their 1995 study involving lesions and training compensatory mechanisms, Wortwein et. al. claim performance isn't, ultimately, degraded). That would put pressure on extreme views about localization and would seem to count as evidence for multiple realizability.

Shelving the connection to those traditional philosophical topics, the point I wish to take out is that survival requires backup plans and behavioral evidence tells us that animals tend to learn them (however we understand them philosophically). So, for some explanation of a behavior, there's usually another explanation for that behavior that covers the organism's back-up plan in case its primary plan doesn't work out. Recognizing this may provide grounds for sidestepping the constraint because it involves recognition of some other

information system that comes to explain the rats' performances when the one thought to explain them is damaged. When it comes to explaining failures in the cue rotation trials, it could be said that the compensatory systems for solving the maze task kicks in or overrides the success systems and produces the failures. Of course, the big problem for this approach is that compensatory mechanisms are thought to kick in when the primary system is damaged or off-line, not when there's failure. Those are different situations. However, the possibility of an alternative, compensatory information system could be appealing to someone who wants to hang on to a success hypothesis while maintaining some theory about the failures. It may seem like an available move.

A second reason for avoiding the constraint may come from steadfast focus on the merits of a hypothesis about success. If a hypothesis appears explanatory, there may be less concern about how evidence of task failures impact it. Instead of worrying about whether it meets the constraint, the move is to punt and point out that some *other* information system explains the rats' failures.

To illustrate, its helpful to think about object-dependent theories of perceptual experience from the philosophy of perception, like direct realism or naïve realism. Prima facie, these theories have the good feature of explaining the justificatory role of experience.[31] My experience of *that* coffee mug on the table justifies my belief that the mug is on the table because I must be visually acquainted with the circumstances that make my belief true in order to have the

[31] See, for instance, Campbell and Cassam 2014 and Brewer 2019.

experience.[32] I must be visually acquainted with the mind-independent mug object on the mind-independent table. However, this good-making feature of the theories precludes the possibility that I could have a perceptual error involving a hallucination of the mug. It cuts off my ability to have an experience as though the mug were on the table even though there is no mug or table at all (maybe I took a pill). Given that these types of perceptual errors could really happen, how could an explanation of perceptual experience that tells us those experiences are object dependent account for them? It seems they cannot, and yet the theories hang around. Their prima facie good features lend them enough explanatory inertia. The standard route for explaining the errors is to adopt a position called "disjunctivism", which states that purported perceptual errors like hallucinations or illusions are not really *perceptual* errors at all. They are all-together a different type of mental state than perceptions and should not be contrasted with perceptual success in veridical perception.[33]

To be clear, my point is **not** that the philosophical theory of disjunctivism is like avoiding the constraint. My point is that the reason why proponents of object dependent theories of perceptual experience pick it as their theory about purported perceptual errors are similar to the reasons why someone would try to avoid the constraint. The similarities lie in the reasons behind the theory about

[32] I could have a seemingly similar experience without the mug if a façade of that mug was placed on the table, but then the belief would be false.
[33] For general accounts of naïve realism or direct realism see Pautz, 2021 and Niikawa, 2023. For general accounts of disjunctivism see Snowdon 2008 and MacPherson 2014

error they pick, not in the theories themselves. They both find a hypothesis (theory) that seems to explain success but recognize the need to explain failures, and so they defer to a different story to explain those. Of course, what's problematic about this type of reasoning is that it the good making features of the hypotheses are tied to the evidence of success. What the error evidence shows us is that the relevant system works differently in the circumstances leading to error.

But whatever someone's reasons are for avoiding the constraint (even if they are bad reasons because they ignore important evidence), trying to sidestep the constraint is deeply problematic and constitutes a 'dead end' for hypotheses. As I will argue, it involves violating deep methodological and ontological commitments of the scientific paradigms they are supposed to contribute to. In doing so, I will have demonstrated that explanations of success in the Morris water maze do not just *seem* to face a constraint, but that they, in fact, do. They are 'locked in' and must satisfy that constraint in order to count as good explanations. There is no side-stepping it.

I start with a basic conceptual point. The information system(s) a hypothesis uses to explain success is either the same one(s) that explains the task failures or it is a different one(s). There are no other options. Were Means, Alexander, and O'Neal's (1992) to offer a hypothesis about the task failures, they would have to point to the same odor information systems they used to explain success or to a different one. Those are the only two options with respect to explaining the failures.

To point to the same one is to meet the constraint, so we can forget about that for now. The challenge for picking a different one is to explain how the separate information system makes any difference to the rats' performances in the cue rotation trials. That information system is supposed to 'kick in' or 'win out' over the success-relevant information system, but how? What is it about the cue rotations that causes or makes a difference to the rats such that a completely different information system from the one it learned to navigate its prior attempts *now*, suddenly, explains its behavior?

It is worth emphasizing that the challenge does not amount to just finding an information system that could cause/make a difference to the task failures. That would only tell us how some information system *could* make a difference to behaviors like the task failures. It tells us something like "*if this information system were to become a difference maker*, then here are the changes it would make to the rats' swim paths". It does not do the job of explaining why that information system is the one that comes to make a difference to rats' performances when researchers rotate the extramaze cues. It is like my mechanic's lazy explanation that a clown-horn is producing the noises coming from my car. I can understand the mechanic's explanation and even agree that a clown-horn under the hood *would* produce those noises. But that doesn't tell me anything about how a clown-horn ended up in my engine. I am right to wonder *how* the clown-horn ended up there or whether the mechanic really even looked in the first place. Absent any answers, the explanation just identifies a possible

source for the noises. It does not tell me how that actually turns out to be the difference maker to my car's horrible noises. In the same way, a hypothesis that only tells us about a possible difference maker to the task failures does not tell us how that system comes to actually make a difference to those performances. That is the important thing that needs to be explained, and explaining it constitutes the challenge.

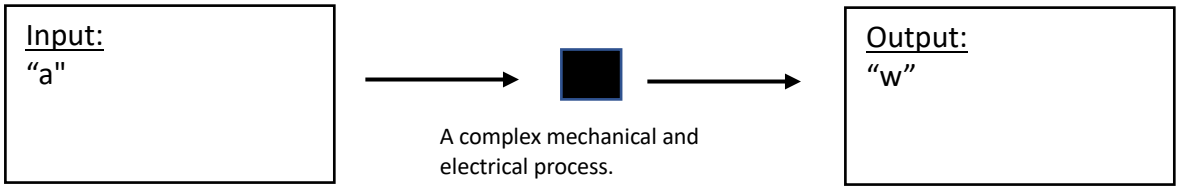
Failure to explain this spells trouble for a hypothesis about the task failures. It shows us that it fails as an explanation of those failures. It falls short of what a *good* explanation should do. To its credit, it does gives us part of an explanation. It tells us how some information system *could* make a difference to the failures. However, it does not explain how that information system comes to make a difference to the performances instead of the success-relevant systems in cue rotation trials.

Hypotheses that point to the same, success-relevant information system have this covered because it is cooked into the hypotheses. In virtue of being hypotheses about success, the functions of those information systems were either learned through reinforcement during training or are innate adaptive traits, and so there is a learned or adaptive mechanism responsible for why the relevant system makes a difference to the performances in success or failure. This is the part of the explanation that's missing if we pick a separate information system. This approach does not explain how the information system that is supposed to make a difference to those failures comes to actually make a difference in the cue rotation

trials. As a result, it does not really explain how the cue rotations make any difference to the task failures. To sum, the explanatory connection between the cue rotations and the information systems that are supposed to make a difference to the task failures is missing, and therefore so is the explanatory connection between the cue rotations and the task failures. The hypothesis does not *really* explain the failures.

To see what I mean, it's helpful to consider a model that schematizes inputs and outputs to two types of performance and the processes that may connect them. Here's how one would look for a cipher machine that turns English sentences into codes letter by letter (Fig.6).

Fig. 6. A scheme for a cipher machine that turns English sentences into code letter by letter.

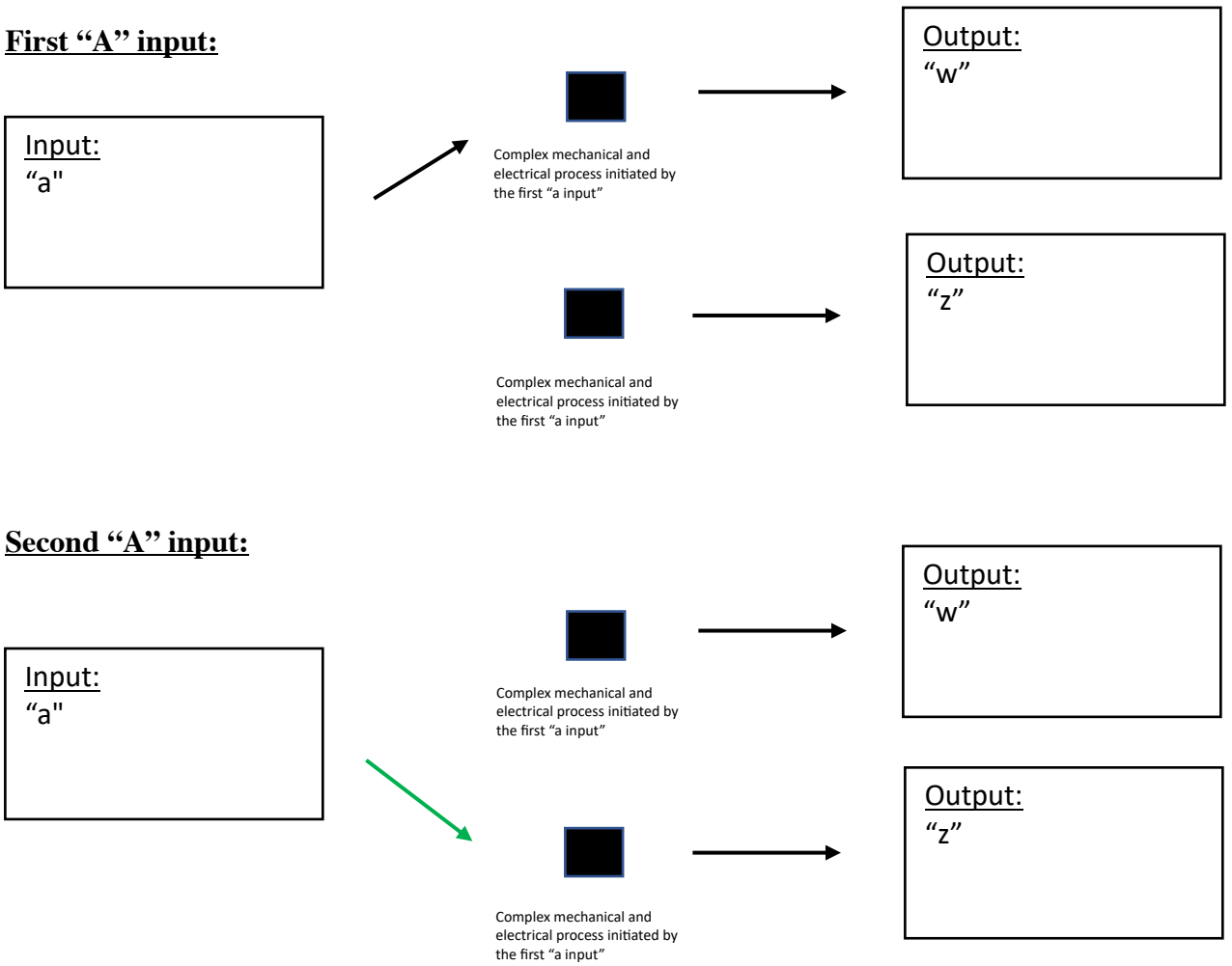


According to the model, when I type the letter "a" into the machine, it prints the letter "w". But it doesn't *just* print out the letter "w". A complex mechanical or electrical process takes place within the machine that mediates between the input and output letters. The details of that process are not so important here, so we can represent them with a black box. The important point is that, according to the model, the input is transformed and manipulated to produce a specific output; it's

not like pushing my hand into the sand and leaving an imprint.

Now suppose the machine is designed according to a rule: the same cipher process never repeats twice in a row. So, when you push the letter "a" twice, even though you hit the same input, a different process kicks in over the old one to cipher the letter. So, if we start typing "aardvarks are nocturnal" we don't get a repeat process with the second "a" input. A different process kicks in and operates instead of the old process to produce the letter "z". Here's what the model would look like with two different boxes for the two different "a" processes.

Fig. 7. A scheme for a slightly more complex machine.



The model becomes more complex. There's a *second* mechanical process, represented by another box, that kicks in to mediate between the input and the new output.

The complexity is appreciated in ciphers because it makes the cipher harder to break. Someone decoding the machine must understand the mechanical process from "a" to "w" *and* the process from "a" to "w". But that's not the only thing adding to its complexity. This is an important point: it's not enough to just understand the processes represented by the boxes. One must also understand the design principle and the mechanisms that led to the second process kicking in over the first one in order to break the code.[34] Absent understanding of this, there is no reason to think outputs like "wz" are ciphers for repeat letters like "aa"; one would probably just default to thinking the same process engages for the same input.

Accounting for this is part of the important contribution of Alan Turing and Gordon Welchman's Turing-Welchman Bombe machine that decoded Nazi Enigma ciphers. The Enigma machine used hundreds of complex mechanical and electrical processes to transform inputs and outputs. But understanding each of these processes was not enough to understand any ciphered messages. That's because the order and arrangement of those processes were highly variable. They could be re-organized or rearranged by adjusting the machine settings. Adjusting

[34] You could imagine representing this with another black box between the input box and the green arrow.

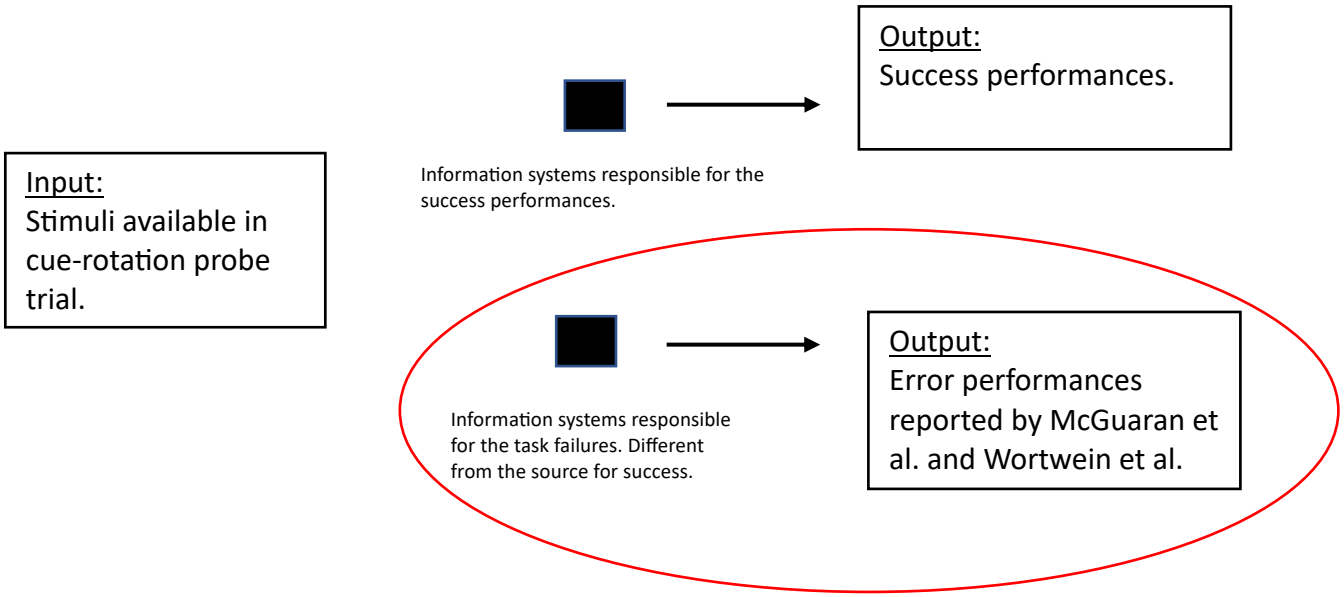
this would change which processes were selected to mediate between the input and output, just like the rule in the example above changed which process was selected to mediate between the input and output. The Nazis coders changed these settings each day to scramble the machine from the day before. The Bombe machine would work each morning to identify the settings the Enigma machines were set up to each day.[35] They would set duplicate machines up, adjust them to the settings, and then decode the cypher.

A hypothesis about task failure is like a hypothesis a decoder gives for the processes in the cipher machine in that they both attempt to explain how a process, represented in the models by boxes, connects or mediates between inputs and outputs. In a hypothesis about failure, the box represents how the rats' inner information systems mediate between the input stimuli available in cue rotation trials and the output task failures. The point I want to make is that failure to meet the challenge reveals to us that a hypothesis fails to do this. Here is why. The hypothesis only fills in the righthand side of the input-output model (circled in red). It only identifies an information system that could stand in for the black-box and produce the task failures. It fails to fill in the lefthand side because it fails to explain how the stimuli in the cue rotation trials lead to that information system becoming the one that 'wins out' or 'kicks in' to guide behavior in those trials. To use William James's helpful terminology, it only explains *a brain path* from

[35] For more on this, see the National Museum of Computing's exhibit on the bombe (https://www.tnmoc.org/bombe) and John Harper's reconstruction of the Bombe (https://www.bombe.org.uk/).

information system to the task failure. It does not explain why that brain path 'kicks in' or 'wins out' over the brain path that leads from the success -relevant system to success performance. As a result, it does not really explain how the cue rotations result in the failure performances. This leaves a researcher where they started with respect to explaining success because it leaves the task failures unexplained.

Fig. 8. Scheme of explanatory links. The explanatory link is missing between the input and black box. There is no answer to the question "why does that black box mechanism activate and not the other?".



To make this point with an example of a hypothesis for successful navigation in hand, let's turn back to Means, Alexander, and Oneil's hypothesis that rats follow odor trails left during training. Remember that the odor trails won't help them explain the failures since there (most likely) wouldn't be any that correspond to the rats' swim paths in the cue rotation trials. To 'hang on' to their hypothesis they would need to claim that some separate information system

explains the failures, and they would need to explain how that system comes to make a difference to the rats' performances in the cue rotation trials. Suppose they say the rat uses visual information instead of odor information in the cue rotation trials, and that explains why they fail at the task. That would explain the connection between visual information and the task failures, but it would not explain why the visual information comes to guide rat behavior instead of the odor information when researchers rotate the extramaze cues. We should ask, "why is the rat, all of a sudden guided by vision, instead of smell?" and "what is it about cue rotation, in particular, that would cause that shift?" If they cannot explain this, then the task failures go unexplained.

To summarize so far, avoiding the constraint involves identifying a separate information system from the one that explains success. A challenge for this option is to explain why that system guides performance in the cue rotation trials. Failure to meet the challenge means a hypothesis about the task failures cannot be an explanation for the failures. It falls short because it doesn't explain the connection between the input stimuli, the inner information system that makes a difference to the performance, and the performance. This leaves a researcher where they started: without any way of navigating around the constraint.

Now, I will argue that there is a deep problem for any attempt to overcome this challenge: trying to explain how a separate information system comes to explain the task failures involves pointing to information system operations that

are not learned or adaptive. Because of this, I regard avoiding the constraint as a dead end.

Understanding why those operations would not be learned operations starts with understanding that they would be novel operations. It would be a new trick. The cue rotation trial is the first time a rat attempts the maze task with the rotated extramaze cues, and so it is the first time the operation would be said to mediate between the rotated cues and failure. In the rat's earlier attempts at the task, it either had no information about the maze, so it explored, or the success-relevant information system guided the rats swim path. The cue rotation trial is the first time the other system kicks in or wins out. It's the first time that brain path has connected the cue rotations to behavior.

That means it cannot be a learned process because learning requires repetition. To say that a brain path or system operation is learned is to say that it activates as a consequence of prior activation and reinforcement. The success-relevant operations are a good example (a literal, textbook example). It led to the rats' success in earlier attempts at the task, and, as a consequence, it repeated when the rat was prompted with the task again. A new, novel response is not a repeated response, and so it cannot be a learned response. To be clear, it could, in the future, be learned, but it isn't learned the first time it pops up. It follows that the novel process that leads to the failure-specific information system guiding behavior could not have been a learned process.

In addition, the process would not be adaptive. If it were not for protocol that instructed researchers to rescue floundering rats from the maze after 60s, the rats would likely perish. But even if they didn't they would find the hidden platform after a massive and frantic energy expenditure. Besides, it is not the new information system that helps them find the platform, it is the explore and search systems that activate the first training trial. So, the new information system only seems to hurt the rat, while other learning systems are responsible for it finding the platform again. If I swim to what I think is the location of a distant buoy only to discover I am wrong, I wouldn't say my intel on the buoy helped me find the shore again. What does is a new search strategy that kicks in after I realize my mistake.[36] Either result would be bad for the rat. The process that produced this result would then be one that reduces the organism's ability to survive in its environment or reproduce, by leading to its death or a big energy expenditure. It follows that the process is not an adaptive process, because it does not help the organism survive in its environment. Borrowing some language from Dretske (1986), getting to the mistaken platform location is not a need of the organism. This is, in part, why such performances are counted as failures by researchers.

So, avoiding the constraint would involve explaining how an information system operation that is not learned and not adaptive guides behavior in all of the rats observed in the experiments by McGauran et al. and Wortwein et al. This should leave researchers scratching their heads. There are no natural, science

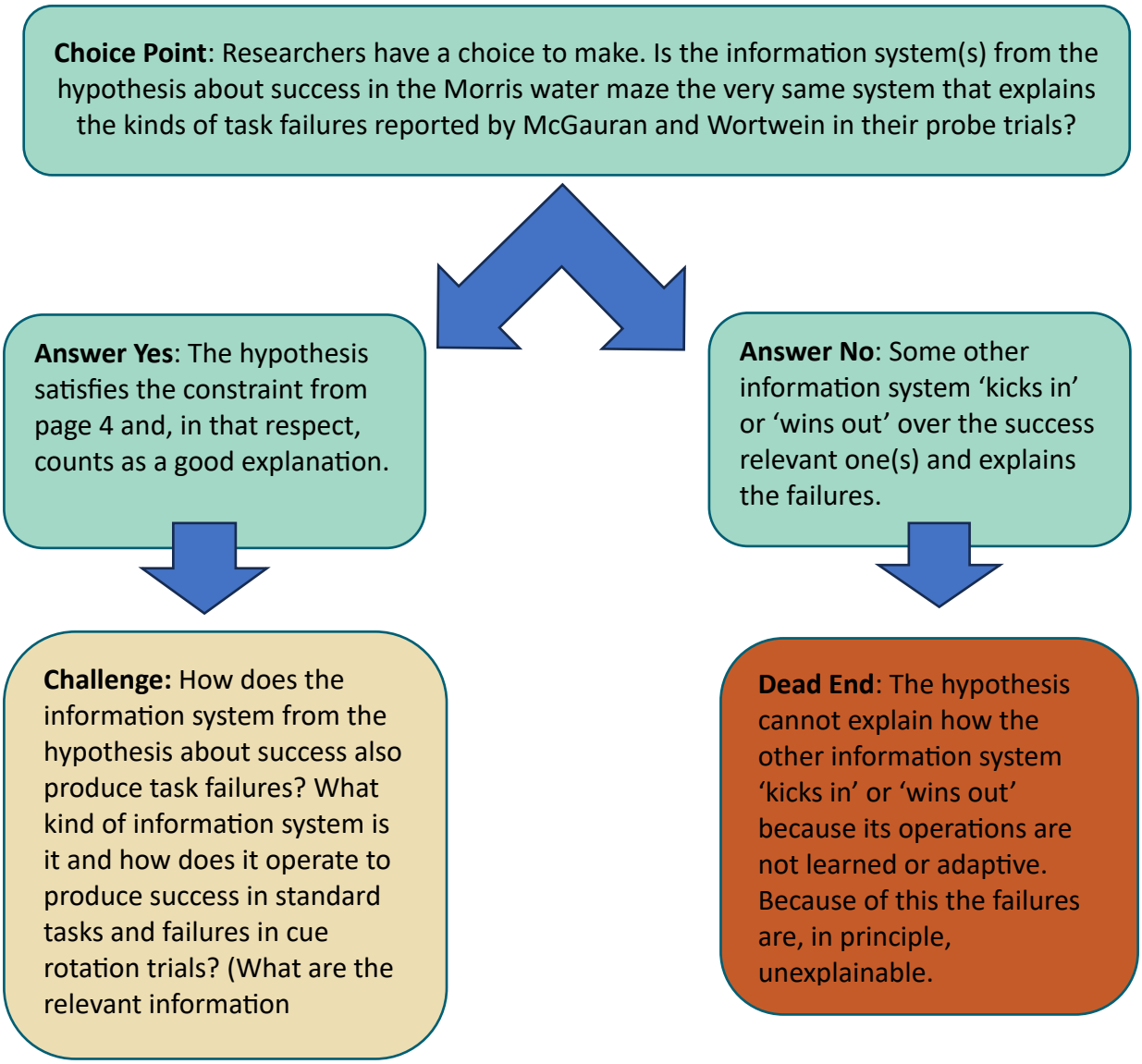
[36] For accounts of the search strategy post mistake, see Wortwein (1995).

friendly mechanisms for explaining how traits or functions persist across organisms outside of learning or adaptation, and so there would seem to be no explanation for how this process persists in all of these rats. Because of this, I regard this first option as a dead end.

Conclusion

Understanding the Morris water maze performances described in Chapters 1 and 2 reveals something of a choice point for hypotheses about task success. In this chapter, I covered the territory associated with answering "no" and following the right-hand side of the diagram (**Fig. E)**. I argued that this option yields an explanatory dead end. Taking the task failures as evidence that some other information system (distinct from the one which explains success) is explanatory of the task failures leaves the failures unexplainable in principle. That is because it requires an account of how that system, whose operations are neither learned or adaptive, comes to make a difference to behavior in cue rotation trials instead of those learned during training. Since this cannot be given without violating the constraints of the relevant scientific paradigms, explanations stuck in this territory would forfeit their ability to explain the task failures performances reported by Wortwein, McGauran, and others. This is too much of a conceit for these paradigms and so I count this option as dead end (Fn 5 on page four offers more information on the importance of these studies).

Fig.9. Choice point for researchers.



In the next chapter, I cover the territory associated with answering "yes" and following the left-hand side of the diagram. On this option, explanations of success at the Morris water maze task all face a constraint. They must identify information systems that also explain the kind of task failures reported by

McGauran and Wortwein. A researcher offering a hypothesis should accept that McGauran and Wortwein have given an important update as to how those information systems work. As it turns out:. I will argue in the next chapter that keeping in line with this constraint imposes something of a tough explanatory challenge. How does the same information system make a difference to two different kinds of behaviors? I will argue that the concepts and strategies we (philosophers and scientists) employ to answer that question that question involving answering tough philosophical challenges.

Chapter 3: Dislocation, Malfunction, and Misrepresentation

The ultimate conclusion of Chapter 2 is that visual cue manipulation studies like McGauran and Wortwein's impose a constraint on explanations of how rats succeed at Morris water maze tasks: **the task-relevant information system(s) from the explanation must also make a difference to the kinds of systematic task failures described by Wortwein and McGauran.** Failure to meet this constraint implies that a hypothesis about how rats succeed can only identify difference makers to task success. I argued that this goes against the way researchers interpret the cue manipulation studies and would require researchers to explain the failures by reference to maladaptive, unlearned systems that violate basic explanatory constraints in the biological and cognitive sciences. So, there are empirical and deeper, theoretical reasons to think the hypothesis cannot be explanatory. A genuine explanation of success must satisfy the constraint and make some explanatory contribution by identifying difference makers or causes of the task failures.

In this chapter, I start to explore what could be said to help a hypothesis satisfy the constraint. More specifically, I look at three approaches philosophers, cognitive scientists, and biologists take with respect to explaining failures, and I consider whether they can help researchers understand how the task relevant systems from hypotheses about success could be difference makers to the task failures reported by McGauran and Wortwein.

I will argue that, while the approaches are good theoretical tools that enjoy successful application more generally, they are the wrong tools for understanding how the success-relevant systems contribute to the failures reported by McGauran and Wortwein. Each one leaves something important unexplained and, seemingly, unexplainable. The specific thing it leaves unexplained differs from tool to tool. In each case though, using one is like trying to use a hammer to fix a car radiator. It just isn't the right tool for the job. It might appear so at first, maybe because it helped us get other jobs done or because it showed promise early on by helping us open the hood or pry off the radiator cap. But, at some point, the hammer stops being helpful and leaves us with important work undone. It cannot help us unscrew radiator hoses or flush fluids. We need to look for other tools to complete these tasks. Similarly, the inadequacy of the three approaches isn't the end of the story. They do not exhaust what could be said. They point to the need for other tools, and in Chapter 4 I continue to explore what else could be said to help researchers understand how the task-relevant systems from an explanation of success make a difference to the task failures. I start to sketch a way of thinking about non-conceptual representations.

This chapter consists of four sections. Sections 1 -3 follow the same format: I describe one of the approaches to explaining failure listed below, some good cases where it really does explain how a system from a success explanation makes a difference to failures, then I argue that the task failures from Chapter 2 are bad cases and the tool leaves the failed performance unexplained and,

seemingly, unexplainable. In section 4, I reflect on what that means for the hypotheses described in Chapter 1 because it amounts to a puzzle: I argued in Chapter 2 that a good explanation of success performances must mention task-relevant systems that also make a difference to the failures described by McGauran and Wortwein. I argue in this chapter that three dominant approaches to explaining the failures run into challenges, and so it seems there are no straightforward ways of forming good explanations of success at Morris water maze tasks.

Three approaches philosophers, cognitive scientists, and biologists take toward explaining failures:

> **1. The Dislocation Approach:** failures are explained by the same internal operations of the organism that led to success and a change to the environment.
>
> **2. The Malfunction Approach:** failures are explained by malfunctioning systems in the organism (but not malfunctions of representation systems).
>
> **3. The Misrepresentation Approach:** failures are explained by misrepresentations in the organism.

Section 1.

Put yourself in the shoes of a researcher running a cue manipulation study like McGauran or Wortwein's. You watch dozens of rats learn and succeed at the standard Morris water maze task (just as expected based on prior research). Then, you rotate a few visual cues outside the maze and watch things go off the rails. All of the rats fail the next trial and spend the next couple relearning the task. How do you explain that? And, in particular, how did the cue manipulations affect the systems you took to explain success? They made one kind of contribution to the rats' success. How is that affected by the cue manipulation?

Suppose you wanted to take a hard stance on the internal workings of the organisms and say they weren't affected at all. That the success relevant systems (like the motor or sensory systems) operate just as you thought they did in success-they do the same thing and have the same effects. Your explanation of the internal workings of the organisms is the same between success and failure.[37] That means the key difference maker(s) that transform a successful performance into a failed one would all be outside of the organism, in the environment. The changes in the environment dislocate the otherwise successful systems internal to the organism. As such, I will refer to this as the *Dislocation Approach* to failure.

[37] That is not to say that every chemical compound or atom of the organism has to be arranged the way it was in success. That's a little too extreme. What remains the same on this approach is what the researchers *say* contributes to success; it is relative to their explanation. If they say system X does F in success, then X also does F in failure.

To illustrate the core idea, it is helpful to think about a Venus flytrap that wastes its digestive juice on a pebble. Flytraps use tiny hairs on the inside of their mouth-like traps to detect movement and signal to its digestive system that its time to liquify the contents of its trap. To avoid false alarms, like if raindrops fell on the hairs, flytraps use a trick. They wait until a few of its tripwire like hairs are tripped, then close their traps and wait. If no further movement is detected, it decides what its got isn't food and opens the trap to wait for something else. If it detects more movement, it initiates digestion.

It's easy enough to imagine a pebble falling and damaging the trap and its tripwire system, leading to more false-alarm digestions. Or a less cautious flytrap that inherited a costly strategy, like initiating digestion whenever any of its hairs are triggered. But let's put a pin in those examples and further constrain our thinking. Imagine it just rained and the pebble falls into the flytrap without much more force than a fly. It jostles a bit causing more hairs to detect movement and the trap to close around the pebble. Now suppose some leftover dew trickles down the pebble and triggers a few more hairs, bypassing the last part of the false-alarm trick, and initiating digestion.

What explains the attempt at pebble digestion? Two things. The same sequence of tripwire mechanisms that initiates digestion of flies AND the dew trickling off the pebble. Were it not for either, the flytrap would not have attempted to digest the pebble (or anything), and, importantly, were it not for the dew, the trick would have worked and the trap would open to wait for something

else. The dewy conditions are crucial to transforming what would have been an effective use of the false alarm trick into a wasteful use of digestive juice. They dislocate the organism's trick so that its normal, everyday functioning leads it to digesting something it shouldn't.

The dislocation approach identifies a similar recipe for task failure. The hallmark of the approach and the way it meets the constraint from the top of the chapter is by a commitment to leaving an organism's internal systems as they were in success. In explanations of task failures, the internal task-relevant systems make a difference to the failure by doing the same thing they did when the organism succeeded. Motor systems produce the same effects and information systems carry the same information in failure as they do in success. Like with the flytrap, the similar internal system operations are crucial causes of the failure. That's how the approach promises to meet the constraint, by saying the success-relevant system(s) makes a difference to water-maze task failures by doing the same thing they did to contribute to success. The only differences it points to are changes in the environment. Those become the crucial resource for explaining failures, so the approach tends to emphasize changes to the environment.

To be clear, the dislocation approach does not say task failures start with changes to the environment that cause malfunctions, or misrepresentations in the organism. It is not the kind of explanation we would offer if we said a mercury spill caused reductions in neuron plasticity or that a thick fog caused us to see a cow as a horse. Nor does it say failures start with changes to the environment that

result in changes to an organism's information even though they were not caused. Like if we were externalists about content and said swapping a frog's supply of flies with small black dots changed the frog's thoughts about what it was eating. These explanations all involve pointing to a change within the organism that resulted from a change in the environment. They involve pointing to a malfunction or a misrepresentation inside the organism that wasn't there before. As such, they are better classified as Malfunction Approaches or Misrepresentation Approaches, and I discuss them in Sections 2 and 3. The dislocation approach explains failures, in part, by telling us that everything within the organism, including its functions and information, are as they were in success. We are to hold that part of the explanation fixed and are left to look solely at the environment for changes that were relevant to the task failures.

To explain the approach further, I turn to a study done by Vorhees and Willams (2006) where it works well, where it delivers a complete story of how merely changing the environment can turn a good performance into a bad one. Like Wortwein and McGauran, Vorhees and Williamson trained rats on the standard Morris water maze task, then introduced special probe trials. However, instead of rotating visual cues outside the maze, they rotated the hidden safety platform the rat is supposed to swim to. The rats trained with the platform in one place, then, researchers move it in probe trials. The rats are like someone who trained for a race that finishes at the base of campus, but because the finish line was moved on race day, runs much slower than they did in practice. Vorhees and

Williamson used several consecutive platform rotation probe trials to determine rats abilities to learn new routes after mastering old ones. In particular, they wanted to determine whether the new routes to the new, rotated platform location would be like the old routes but with extra steps tacked on to the end or whether they would learn a novel, direct route.[38]

I want to focus on the first attempt at the platform rotation probe trial and set the rest of the experiment results aside. Unsurprisingly, rats swim to the old platform location they had been trained to swim to, and, as a result, fail their first probe trial. They take longer in general, spend longer in the wrong parts of the maze, and take a longer swim path. What explains the failures? Why do the rats take longer or take longer swim paths? The Dislocation Approach shines here; it gives us a complete story. The rats fail at the task because the internal system operations it learned to successfully navigate to the platform operate the same way in the platform rotation trial and take it to the same location it swam to in training. And why wouldn't it? The safety platform is invisible to the rat, so it has no idea it moved when it starts the task. The key to the failure is that researchers moved the location of the platform out of the way of the swim path. They changed the environment so what would have been a successful swim now counts as a failure. The dislocation approach is perfectly adequate for explaining these failures and others like it.

[38] They report an interesting difference between species. Mice learn navigate to the new platform by adding steps to the old routes, while rats learn novel routes.

But does it help us understand the task failures from McGauran and Wortwein's *cue* rotation trials, where researchers leave the location of the platform alone and only move visual cues outside of the maze? I don't think it does. The failures are importantly different. The hidden platform is in the same place it was during training, and the rat swims in the wrong direction. What needs to be explained is the rats new swim path in probe trials, its veering off in the wrong direction. In Vorhees and Williamson's study, they left the visual cues alone and only moved the hidden escape platform. What needs to be explained is totally different; it is how the *same* swim path that led to success during training led to failure in the probe trials.

The dislocation approach takes a hardline perspective on causes or difference makers to the new swim path that are internal to the rat. It says they make the same contribution to the new swim path that they did to the old ones. Whatever mechanisms were said to go into motion, whatever effects they produced, and whatever information carried that led the rat swim directly to the platform in training must also be said to lead it off course in the probe trials. A biologist explaining what caused the rat to veer off and fail would point to the rat and says "Well, there's no use looking for any changes in there. Everything is as it was during training". The only resource it has for pointing to changes that affected the new swim path are external to the rat in its environment, not in the rat.

So, if we used the approach in combination with Means, Alexander, and Oneal 's (1992) theory that rats complete the task by following odor trails left in training, we would hold the internal odor-detecting operations of the organism fixed. We would say the information and motor systems track and follow old odor trails just as they did in success and constitute important causes of the new swim path. But this should raise red flags. The swim path is a *new* swim path. The rats never veered off in that direction during training, so there are no odor trails for them to follow in that direction. Our explanation hasn't identified any difference makers to the new swim path. Worse, is that the difference makers we have identified would predict success at the task. If we really thought the rats were following odor trails in the cue-rotation probe trials, we would be surprised to find out they fail the task.

Because the approach takes such a hardline on the internal systems of the rat, the only other explanatory resources available are in the environment. But those don't lend much help. The environment is highly contrived and controlled. The only change is the rotation of the extramaze visual cues. That's the only difference in the environment that could explain the change in swim path, and it has no effect on the task relevant odor detection systems we've selected in our explanation. Moving the extramaze cues doesn't produce a new odor trail or have an effect on the platform location. So much for the combination of the odor-trail theory and the dislocation approach, the combo fails to identify causes or difference makers to the novel swim path followed in failure.

Combining the approach with other theories of navigation that don't identify the extramaze cue as a difference maker to success runs us into the same problem. If, for example, we pick a theory that says rats use intramaze visual cues or a list of prior body movements that took it to the platform, then we would, again, fail to identify difference makers that put the rat on the new, wrong-direction swim path. If we say the extramaze cues are not registered in success and adopt the dislocation approach, then we are committed to saying they are not registered in task failure either. It follows that there is no reason or cause for the new path in the probe trial. Like Means, Alexander, and Oneal's theory, they predict success. That's part of the reason the extramaze cue rotation studies are so important and have been replicated so frequently. They provide strong evidence that extramaze cues are an important part of any explanation of success at the Morris water maze task.

So how does a theory that highlights the role of extramaze cues do when paired with the dislocation approach? One that says rats succeed at the task by using information about the location of the platform relative to the location of the cue. Paired with the dislocation approach, we would say the rats swim in the wrong direction because they uses the same systems to collect the same information about the extramaze cues and relative platform location and that, because of the cue rotation, we should expect the rat to swim in the direction of the rotated cues. Doesn't this appear to identify relevant difference makers that explain the failure?

It does, but it fails to make good on the hallmark of the Dislocation Approach. It fails to hold the internal operations of the rats fixed and involves saying something different about the internal goings-on of the rats between success and failure. In particular, it involves saying the information used by rats is true/accurate in success and false/inaccurate in failure. This *has* to be said. The rats must use the same pieces of information about the relationship between the extramaze cue and the platform it learned during training or else we would violate the core idea behind the approach by saying they used different information. If it learned to swim to the platform by thinking something like "If I am facing the extramaze cue, the direction of the platform is located at 3'oclock relative to that cue'", then they must also think that when they fail. But keeping the same pieces of information requires us to say there is another difference. Not in the information's content, but in its relationship to how things really are in the environment. It yields a difference in its veridicality, accuracy, or 'truthiness'. The information about the relationship between extramaze cue and the location of the platform the rats learned during training matched up with the actual relationship. They reflected how things really were, which is why the rat was able to navigate to the platform. The cue rotations in the probe trial dislocated the extramaze cue and altered the relationship between it and the hidden platform. It is no longer at 3'oclock relative to facing the cue. The rats, not privy to the manipulation, hang on to the same information about the relation, but the

information is now inaccurate or false, which seems to explain why they veer off in the wrong direction.

It's like if you regularly walk to the goldfish pond in Pogonip from all sorts of entrances around the park and learned to do this by using information about the pond's relationship to a far-off, distal cue outside the park like the clocktower downtown. Whenever you want to get your bearings and determine which direction to walk in, you scan the horizon for the tower, remember that if you are facing the tower the pond is at 3 o'clock relative to your facing, and head in that direction. Now imagine that the city of Santa Cruz moves the tower ten miles south while you're away on vacation. You come back without realizing the difference and use the same strategy to navigate to the pond from a new start location. You look for the tower, face it, and head off in the direction of an hour-hand pointed at 3'oclock. After walking in that direction for a few hours and realizing you are lost, you pull out your phone GPS and see you are way off course. You think to yourself "That's weird. I did exactly as I always do. I used the same navigation trick". When you later learn about the moved watchtower you realize there was something different about the trick that steered you in the wrong direction. The information you used was the same in content but different in its relationship to the world. It was inaccurate or false and that, coupled with your use of it anyway, was the reason you ended up in the wrong place. Or, maybe you are a content externalist and think changing the location of the tower actually changes the content of the information you use. Either way, you are

forced to say that something changed between your successful and unsuccessful treks to the koi pond.

Because of this, attempts to explain failure in terms of information about extramaze cues while holding the internal operations of rats the same as they were in success are doomed to fail. They must say something different about the effects the extramaze cue has on the rats' information. They must say there is a spontaneous change in information in task failures, which is an obvious offense to the approach and looks more like the approach argued against in Chapter 2. Or they must say the information is inaccurate or false, which looks more like the Misrepresentation Approach I will characterize in Section 3 (and that has its own set of problems). Attempts that circumvent this by ignoring the effects of the extramaze cue manipulations can hold the internal operations fixed, but they do so at the expense of failing to identify difference makers to the novel swim path in task failure. For these reasons, the Dislocation Approach does not help explain the task failures described by Wortwein and McGauran.

Section 2

It seems that explaining the rats' failures involves saying something about how the cue manipulations affect the success relevant systems inside the rats. It involves saying something about how they operate one way in success and are so affected by the manipulations that they operate another way in failure. In this section, I consider whether thinking about the manipulations as causing a

malfunction of those systems can help us understand how they contribute to the failures.

The Malfunction Approach is fundamentally different from the Dislocation Approach in that it explains task failure by reference to changes inside the organism. It says the task relevant systems operate differently between success and failure. They operate as they should in success and malfunction in failure. What makes the difference between good and bad performances at a task is whether the relevant, internal systems of an organism function appropriately or not.

The approach is familiar enough from applications in everyday life. What explains why my phone application closed? Or why my car will not start? Because, we find out, some part or subsystem of those machines are not doing what they are supposed to. They are not doing what they were *designed* to do. When my alternator stops taking energy from the car's drivetrain to re-charge the car battery, the actual operations of the alternator come apart from its designed operations. As a consequence, my car doesn't start.

Application of the approach to natural or biological systems is made complicated by the fact that it involves claims about what those systems are supposed to do. Saying that something *isn't* doing what it is supposed to involves saying something else about what it *is* supposed to do. However, natural scientists and philosophers sympathetic to their explanatory goals cannot say the systems were designed to operate a certain way like the car alternator because it involves

positing something *unnatural,* like a God-like designer who designed the natural world to operate in certain ways. Some philosophers treat this as a problem for claims about what natural systems are supposed to do. Kant (1789), for instance, think it is evidence that natural scientists are not licensed to make claims like that at all, while Aquinas (1269) thinks their necessity in areas that we now think of as natural science is proof of the existence of a designer God.[39]

So, if biological systems are supposed to do something or, to use the technical lingo, have functions, then they must have them in virtue of something besides being designed to have one. Accounts of how they have them usually point to the history of the system. They tell us selection or training for a system operation took place sometime in the system's past, and *that* part of the system's history is sufficient for the operation to be the system's function. The idea, in rough outline, is that biological system operations were selected for or learned via training. Systems perform their function when they do what they were selected for or learned to do. They malfunction when they fail to do what they were selected for or learned.

Human hearts that pump blood do what they are supposed to because they do the same thing the hearts of those humans' (recent) ancestors did that led to their selection. Those ancestors' hearts pumped blood, and their pumping blood was a cause of the ancestors' reproductive fitness which led to copies of the heart-

[39] See, for example, in Kant, Section IV 1st introduction and in Aquinas, Question 2 Article 3. For discussion see Plantinga, 1993 (ch.11) and Garson, 2007 and 2011.

system in their children. A human heart that does not pump blood, on the other hand, does not do what it was selected for and so it malfunctions. The account of how hearts come to have functions doesn't need to point to a designer of a heart to assign it a function. It can point to some aspect of the heart's history. This is important to note at the outset because it constrains what we can reasonably assign functions, and because they direct our attention to the appropriate grounds for claims about malfunctions.

To be sure, there are other ways of using the word "function" grounded in other facts about biological systems. Claims about functions are used to accomplish a wide range of desiridata in philosophy and science, like explaining the presence of a system operation in a single organism, the distribution of an operation across a population of organisms, the difference between essential features and accidents, assigning representational contents, and more. They are not always aimed at accounting for how it is a system can malfunction-they are not always aimed at explaining *normativity*. Cummins (1975), for instance, argues that functional ascriptions are claims about a subsystem's contribution to the larger, overall system(s) it is a part of. Determining the function of a system has to do with how we decompose or break an organism down into subsystems and think about the current contributions of those subsystems to the goal or health of the larger systems it is a part of.[40] A human heart functions to pump blood because that is what contributes to the goals or health of the human it is a part of.

[40] Or whatever it is about the larger system we are interested in.

The heart's past doesn't matter at all. It doesn't matter if it was selected for or created in a person's chest by spontaneous miracle.

Cummins's theory is meant to explain how functional claims support scientific analysis of the present goings-on of biological systems. It is not meant to explain the normativity of those systems.[41] If we try to explain the irregular activity of a frog, who can't jump as high or as often as other frogs due to injuries to its legs, we can still think of a function or role the legs play in those activities. Even if we would say that role isn't the same one it played in ancestor frogs or other frogs, there's a sense in which the legs still make contributions; they still do something. Sometimes philosophers and scientists use "function" in this way so it is a matter of what a system is presently doing. But when they use it this way, it is difficult to see how those systems could malfunction. Since, if it is not presently doing something, then that is not part of its function. So, it doesn't malfunction when it fails to do it. That's just to say that some of the ways philosophers and scientists talk about functions and malfunctions are not specifically aimed at capturing how malfunction is possible. Assigning a function to a system does not automatically mean it is amenable to malfunction or that failures of the organism it is a part of can be handled by the Malfunction Approach. This will be important to keep in mind later, as we try to find malfunctions that can explain the failures.

[41] For critical discussion of whether Cummins-like theories can explain malfunction see Godfrey Smith 1993 (p.7), Neander, 1991, and Garson, 2011

Once a malfunction is found, they can be traced to changes in the environment or mutations/disease within an organism. To illustrate, consider ocean bacteria that rely on compass-like organelle that 'point' in the direction of the strongest local magnetic field. Usually, they point down in the direction of the Earth's core to the oxygen depleted water the organism needs to survive. Suppose we dove down with high-powered microscopes and observed bacteria swimming up in the wrong direction toward toxic, oxygen rich water. Concerned by this, we zoom in on their compass organelle and see them working correctly, pointing down toward the oxygen deplete water. Phew, no issues there. Now that we know those are good, we search with our microscopes and find malfunctions in the bacteria' locomotion systems. They are supposed to have fin-like cells that flap to turn the bacteria in the compass direction, just like their ancestors. However, the cells don't appear strong enough to flap fully and turn the bacteria in the right direction. They malfunction. We can trace the malfunction to one of two kinds of causes. Causes in the environment, like damage to a swimming fin from predators or sediment in overwhelming ocean currents. Or internal causes like mutation or disease.[42] The malfunction approach makes no exceptions based on the cause of a malfunction.

[42] One problem for identifying internal causes like disease or mutations as a source of malfunctions is that a disease or mutated system may lack the physical capacities for having the relevant function in the first place. For instance, saying a particular magnetosome's locomotion fin flapping system lacks the capacity to turn it in the direction the compass organelle is pointing might disqualify the fin-flapping system from having the function of turning the bacteria in the direction the compass is pointing. As a consequence, we might say the locomotion system is not the same kind of system found in other bacteria that can and do turn the organism in the direction the compass points (we might say that if we type systems by their function, for instance). The

There is one important exception to the Malfunction Approach, one way of explaining error or failure by pointing to a malfunction that does not count as application of the Malfunction Approach. It involves pointing to malfunction of a representational system. Suppose, for instance, that I want to explain why I mistook Franco for being angry by pointing to a malfunction in my visual system. I say that I suffered something like a visual illusion because my fear of upsetting Franco led me to see their smile as a down-turned frown. I say that my visual system was supposed to deliver an accurate representation of Franco, but it didn't because of my fear of Franco affected what I saw.[43] While it is true my explanation involves attribution of a malfunction, the important function it is tethered to is a function *to represent*. And assigning representational functions involves a distinct set of explanatory tools and challenges. If we say the visual system has the function to represent, part of what we are saying is that the visual system is intentional or has content. That it is *about* something else the way a painting or sentence is. Just like we may face questions about what a painting or sentence is about, we might face questions regarding what the visual system is about and why. A biologist explaining the malfunction of a heart doesn't use those tools or face those challenges. As such, I do not categorize explanations that point to malfunctions of representational or information systems under the

---

latter systems were selected for and so have the relevant capacity and function, while the former were selected against and do not. Since they do not have the relevant capacity and function, they do not malfunction. They are not systems that have the ability to do X and fail to; they are systems that were selected against because they lack the ability to do X in the first place. See, for instance, Davies 2000 and Sullivan-Bissett 2017.

[43] For further discussion of this example, see Siegel, 2010 and 2017

Malfunction Approach. They are a different animal with different problems, and I categorize them as Misrepresentation Approaches in Section 3.

The Malfunction Approach works really well for explaining one-off behavioral anomalies. Morris water maze studies often involve dozens of animals in up to a dozen trials, and thing's don't always go as planned. Take a look at the data provided by Gehrin et al (2015) from their study measuring swim paths of rats in Morris water mazes (Fig.10). Graph A shows outlier behavior in trials 10 and 11 marked by crosses. Those rats had a remarkably longer swim latency (total seconds spent swimming to the platform) compared to other rats. Gehrin et al don't explain why they took so much longer, but it wouldn't be surprising to learn they suffered injuries, were sick, or unmotivated. Malfunctions in the arms or legs due to damage from mishandling or a fall, cardiovascular disease due to a bad diet or stress, or even random genetic mutations that make the rat worse at swimming are all relevant difference makers to one-off failures like these. These are credible, relevant difference makers to that kind of behavior.

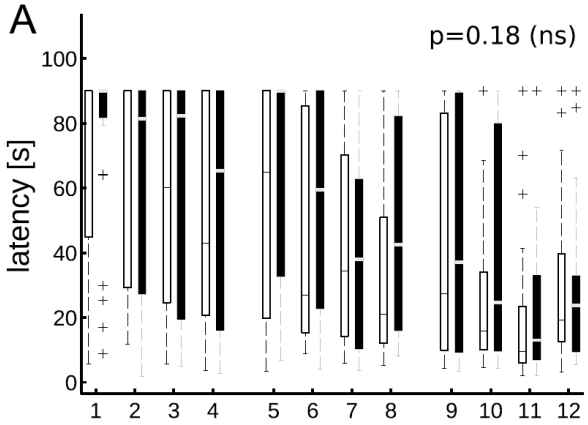


Fig. 10. Latency metrics (y axis) for two groups of twelve rats over a set of 12 trials (x axis). White boxes are control group. Black boxes underwent variable protocol (stress). Boxes represent first, second, and third quartiles with median as band. Outliers marked with crosses.

The Malfunction Approach can also explain systematic failures. For example, in a study measuring the effects of MDMA on spatial memory, Cavieresa et al (2010) injected rats with different amounts of the drug and found that it diminished their abilities to learn new maze tasks. Rats given low doses did worse on tasks than rats given Saline for control, and rats given high doses did even worse than the low dose rats. Cavieresa et al (2010) explain that MDMA has a toxic affect and decreases the long-term potentiation of neuron cells in the hippocampus. The affected cells are too weak to send the strong signals they are supposed to send, and this malfunction results in failures at the task.[44]

The problem with using the Malfunction Approach to explain task failures from cue rotation studies like McGauran's and Wortwein's is that there do not seem to be any malfunctions in the task relevant systems. First, consider we are limited in what we can say about the sources of any malfunction. All the rats fail in the probe trial, so failure cannot be traced to anything like one-off diseases, injury, or lack of motivation like in Gehrin et al's study (Fig A). The source of the malfunction, whatever it is, must produce malfunctions in *all* of the rats, which means it must be part of the probe trial task. But the probe trial doesn't involve direct interventions to the task-relevant systems like studies involving stress protocols, pharmaceutical interventions, or lesions. Caviresa et al's MDMA study is a good example; they give the rats MDMA before the probe trials. The cue

[44] Similar experiments are reported in Vorhees and Williams 2006 and Brandeis, Brandys, and Yehuda 1989

rotation probe trials are relatively gentle in comparison. The only change is the rotation of extramaze visual cues. Any malfunction in the probe trial must be traced to this.

Finding malfunctions is made even more difficult by the rats' behavior in the probe trials. They don't freeze, tremor with fear, writhe in pain, stumble, bump into the walls, or seem, in any way, uninterested or incapable of the task. They B-line to the place they think the platform is, in pretty much the same way they did in training. This constrains what we can say has gone wrong in the probe trial. Pointing to malfunctions in systems like the locomotion, cardiovascular, or circulatory systems just won't do.

The best place to look for malfunctions would be in task relevant systems that function to carry information. But, before considering this option, some disentangling is in order. So far, I have used the word "information" in the way working scientists do. To describe a wide variety of phenomena, from full-blown conscious perceptions to the activity of specialized neurons that only 'fire' in the presence of something super specific like a red square in the center of my visual field. As a consequence, some of the things I call "information functions" turn out to be representational functions too. Let's shelve the representational ones, because, as I noted earlier (p.9), appealing to them involves using concepts and tools associated with the Representation Approach I will discuss in Section 3. The information functions I have in mind are more like the ones we come to assign to the blinking blue lights on a pair of Bluetooth headphones. When I see the lights,

I think "ah, my headphones are pairing" because I think it is part of the blinking light's job to indicate the headphones are in pairing mode. I come to think of them as having that job because I see, over and over, that when I put them in pairing mode, the blue lights start blinking. That's really all there is to the sorts of information functions I have in mind, normal covariance or lawlike causal relationships. These align with the sorts of functions that Stampe (1977) and Dretske (1981) use to describe and think about the content of information states.

Non-representational information functions like this can be jeopardized without damaging or physically intervening in the system carrying the information. Situations involving confounders or high-noise are good examples; just imagine a specialized 'cow-detecting' system that lights up an LED in the presence of a spotted horse or a horse standing in fog. No tampering with the system is needed, you just need to 'trick' its detectors. That means we can think of malfunctions of information systems as steering the rats in the wrong direction without having a more detrimental, direct impact on the systems responsible for the rats' behavior. So, malfunctions of information functions seem to fit with the details of probe trial. They can be traced to cue rotations, and they don't directly interfere with the rat's health, movement, or motivation. They would be very helpful-perhaps the best-case scenario for the Malfunction Approach. The problem is, again, there don't seem to be any.

To see this, suppose that, like so many researchers studying neural activity of rats during maze tasks, we find evidence of specialized neurons during

training.[45] We find neuron cells that only deviate from their baseline rate of electrical activity to do something exciting (like exhibit complex spike patterns) when very specific conditions are met in the task. Maybe we find evidence of *place cells* that only seem to fire when the rat is in a specific local in the maze, the cell's *place field* (Fig. 11), or *boundary cells* that only seem to fire when the organism nears a boundary wall of the maze (Fig.12). We could infer these cells are carrying information and that they have the function to do so. Note that this would be a very general claim. We would not yet be claiming they have the function to carry information about anything specific. We can try that later. For now, suppose they just have the function to carry *some* information. It's like when we see some telephone infrastructure from a distance and know the poles and wires are meant to carry information, but we don't know if they have more specialized roles to carry information from Verizon or AT&T customers.

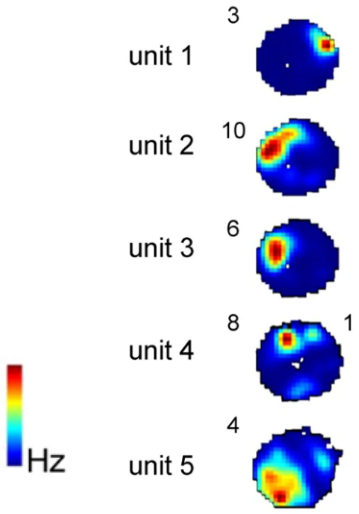


Fig.11. Place fields for five place cell neurons. Place fields characterize the places in an environment where a neuron (or cluster of neurons) exhibited interesting activity. Interesting activity is characterized with yellow, orange, and red. Uninteresting (baseline firing rates) are characterized with blue (peak activity represented in HZ to the left of the circles). If we look at Unit 5, we see that this neuron cell only exhibited interesting neural activity when the rat was in the Southwest quadrant of the maze. Park et. al. 2015.

[45] For examples that illustrate discovery of this sort of evidence, see Ranck Jr. 1973 and O'Keefe, 1976. For more, current examples, see Otto and Gallagher, 1995 and Schimanski et. al, 2013. For reviews and discussion see Reddish, 1999 and Kneirim and Hamilton, 2011.

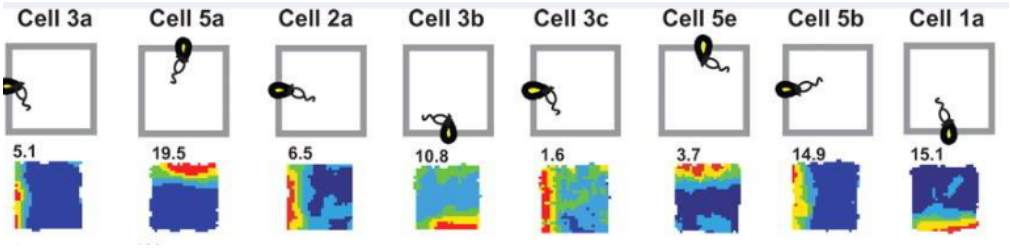


Fig. 12. Boundary cells exhibit interesting activity whenever the organism is near a boundary in its environment. Cell 3a, for example, only exhibits interesting activity when the organism visits the West wall. Lever et. al 2009

So far, so good. Assignment of a general information function seems appropriate. The cells really do seem to have the function to carry information. But does the assignment help us find a malfunction? Can we really explain failure in the maze task by saying the extramaze cue rotation in probe trials caused a malfunction of this general information function?

It's hard to see how it could. The neurons will either exhibit their interesting activity in the probe trial or they won't. If they do, then they are carrying information. The neurons are carrying information that the special circumstances they have been trained to detect are occurring. But, even if they do not exhibit their interesting activity, they still seem to carry information. They just carry the different piece of information that the special circumstances they have been trained to detect are not occurring. Imagine an LED light that is programed to only turn on when someone is outside your office. Even when the light is off, it tells you something: no one is outside your office. The cue manipulation doesn't

seem to upset or breakdown any general information functions, and so assigning such functions doesn't seem to help us find a malfunction.[46]

Looking for a malfunction of the most basic information function, the function to carry information, will not help us because we will not find one. But what if we got more specific about the information function? We could say the neurons we observed during training had the function to carry information that is accurate or true. Some philosophers think of vision and early vision processes as having a function like that (Fodor, 1985. P.4). The question is whether thinking of the neurons as being like vision can help us find a malfunction to explain the task failures. The idea is that the cells do their job and carry *accurate* or *true* information during training, but malfunction during the probe trial because they carry inaccurate or false information that doesn't reflect the cue manipulation.

The problem with this is that it just isn't application of the Malfunction Approach. Assigning the function to carry *accurate* or *true* information smuggles representational concepts into our explicitly non-representational approach to explaining the failures. It involves saying the neurons do more than bear a law-like covariance or causal relationship to whatever it carries information about.

---

[46] It's not even clear what would count as evidence of malfunction of a general information function. Thinking about the LED light is helpful for illustrating this. We might discover the battery is dead or the light has been unplugged and think "ah, now the LED is malfunctioning, it no longer carries any information". But is that the right inference? Couldn't we say the LED's 'being off' carries information there is lack of a power source? In which case, it would still carry information and satisfy the function. Likewise, if we found neuron cells that didn't exhibit even baseline electrical activity, we could still think of them as carrying information that the neurons are damaged or have been taken 'offline'. Information, in the sense I've been using it divorced from representations, seems ubiquitous; it is everywhere. For discussions of the ubiquity of information see Ramsey, 2007.

Like that they are about it in ways that involve appeal to representational concepts and tools. Or that they have content that is structured and can be accurate or false. The relationships of interest here, on the Malfunction Approach, are more like the ones between the flashing blue-lights on my headphones and their pairing mode, and we wouldn't say the lights carry accurate information or true information about pairing mode.

What about getting *really* specific about the information the neurons are supposed to carry? Suppose during training we saw a cluster of neurons that only fired when the rats were facing a specific visual landmark so that the platform was in the direction of an hour hand pointed at 3 o'clock. We could reasonably infer the cells have the function to carry very specific information about the platform being located in the direction of an hour hand pointed at 3 o'clock. Then, we would seem to be within our rights to say the cells malfunctioned in the cue rotation trials. We could say the cells are only supposed to fire when the platform is in the direction of 3 o'clock, that it is their job to do so. However, the cue rotation causes them to fire when the platform is located in the direction of 7 o'clock (or wherever it ended up after the rotation). As a consequence, they carry the wrong information. They are supposed to be snobby about their firing and only carry 3 o'clock direction information, but they are carrying 7 o'clock direction information. This would cause the other systems (like locomotion systems) to operate as though the platform were at 3'oclock, which generates the failure. That is a plausible story of how malfunction could produce the failure.

The problem with this lies in our functional assignment, with our claim that the cells have the function to carry specific information regarding the platform's location in the direction of 3 o'clock. Suppose your lab mate says "no, no, no, that is not the cells' function! Sure, they have the function to carry some very specific information, but not *that* specific information. They have the function to carry information that the visual landmark is located dead-ahead, in the direction of an hour hand pointed at 12 o'clock noon. Don't you see that is when the cells fired during training? What's more is they continue to do this in the probe trial, and so they do not malfunction. They are doing their job even in the probe trial!".

You and your lab mate disagree over the function of the neurons. To be clear, you both use the word "function" in the same way. You both think the cell function is determined by what it learned/was selected for. You're not slipping into some of the other was scientists or philosophers, like Cummins, talk about functions. You just disagree over what these cells learned (or what they were trained to do). Sure enough, when you look back at the training data, you find the neurons did fire when the rat faced the landmark, even in the probe trial. That's because the platform's location in the direction of 3'oclock covaried with the rat facing the landmark, so it makes sense the data would reflect that. We should expect a system that detects the latter to also detect the former. And so, our data about the neurons behavior during training provides grounds for both function assignments. There is functional indeterminacy. On one of those functional

assignments, we can reasonably motivate a malfunction. We can say the neurons were specialized to carry information that the platform is in the direction of 3 o'clock, and that they did not carry that information in the probe trial. But there is no reason to prefer that assignment over the other given by your lab mate, on which there is no malfunction in the probe trial. So, even if we get really specific about information functions, we still run into problems finding malfunctions.

It is worth noting that it is our lab mate who makes the alternative functional ascription, not a skeptical philosopher who points to functional assignments working scientists would never *actually* point to (the situation is not a 'philosopher's fiction'). It's not like the functional indeterminacy that results from claiming the neurons actually have the function to detect proximal properties, like light with *n* wavelengths, and that the neurons do not malfunction because they do that job well in the probe trial. Our lab mate's suggestion isn't anything like that, and we cannot dismiss it as a philosopher's fiction. It is what Karen Neander calls *plausible* indeterminacy that scientists would agree on (Neander 1995). The idea being that working scientists really could arrive at this indeterminacy and feel the tension between saying the cells function/malfunction.

It is also worth noting that looking at how the receivers or consumers of the information use it will not help us pick from the two functional ascriptions. The strategy may be helpful (or even essential) for understanding what an

information system is about in other contexts.[47] Observing bees fly to the outskirts of Porter Meadow after they watch a fellow bee's waggle dance can help us determine that the waggle dance meant something like "There are some good flowers along the outskirts of Porter Meadows" or "Hey! Go to the outskirts of Porter Meadows!". But it cannot help us here. Neural activity just doesn't fit the ways we constructively think about receivers/consumers of information. Consider, first, that demarcating or dividing the brain into discrete entities such that one could count as a receiver is not straightforward. Is an individual neuron a receiver? A cluster or group with similar activity? Or should we speak at a higher level about brain structures, like the hippocampus, or brain systems, like the visual system or navigation system? What sort of thing is it we have in mind?

However we decide, we run into another issue. The activity of any of those things (neurons, clusters of neurons, brain structures, or systems) is limited. All they are really capable of doing is exhibiting firing behavior. Some of which is interesting because it deviates from baseline, normal firing behavior. So, our characterization of the activity of receivers must be in terms of this activity. We'll have to say something like, "the receivers of some piece of neural information are those discrete brain entities that fire in close temporal and spatial proximity to the neurons carrying the information".[48] But the rat brain is a busy place. There are

[47] Millikan argues that understanding the proper functions of receivers and senders of a signal are essential to understanding the content of the signal (1984 & 1989). For discussion see Godfrey-Smith 2013.

[48] More successful characterizations of receivers focus on things like the receiver's coordination or mutual benefit with a sender, but those won't work here because it is not clear what is good or what benefits a neuron or cluster of neurons. See, for example, Lewis 1969 or Skyrms 1995.

roughly 200,000,000 neurons firing very frequently (about once a second). It's also interwoven. Any neuron can have close proximity synaptic connections to hundreds or even thousands of other neurons. So, we can expect a lot of interconnected neurons, maybe hundreds or thousands of neurons, each fitting the bill of firing in close spatial and temporal proximity to the neurons carrying the information. What's more is that the characterization of receivers must be similar to the characterization of what senders of other signals are doing. Or of the signal itself. So, this approach to the functional determinacy yields another kind of indeterminacy over what counts as a receiver and why. Here is Godfrey Smith summarizing similar points nicely.

> "[…] surely the brain can be seen as a signaling device. Neurotransmitters transmit signals between neurons, for a start. But whether this kind of activity fits into the sender-receiver configuration […] is not so clear. If we look inside a brain and find a huge network of neurons, each affected by some and affecting others, it appears that any one neuron's firing might be described as either a signal, or the reception of a signal by a receiver, or the sending of another signal, depending on how one divides things up." Godfrey Smith, 2013 p. 52.

One last consideration relative to the indeterminacy involves a call for more studies or trials. The idea is that researchers could solve the indeterminacy

by conducting more studies in which they remove the visual landmark and place the rat in the maze so that the platform is in the direction of 3 o'clock. If the cells fire, then it seems like evidence they have the function to carry information about the platform, not the visual landmark. The problem with this is that neuron cells are known to 'remap' and change their jobs when environmental conditions change. This is a well-known and studied phenomena (Knierim and Hamilton 2011, Jefferies 2011). Study the way neurons behave in one environment and you see patterns. Change the environment enough, and you see different patterns. That's part of what makes the cells so interesting. It's likely that changing the environment by removing a landmark the rat has focused on and obviously depended on in training, in a maze environment that is designed to restrict differences in sensory signal to reduce confounding effects, could change the specialized jobs of those neurons. In a similar maze study, Boston and Kubie (1991) found that manipulating the color of a single visual cue (black to white) was enough to cause neurons to remap. If the jobs change, we cannot use the evidence to corroborate our initial function assignment. Besides, thinking about how the neurons fire absent the extramaze cue obfuscates the role of the extramaze cue. If the rat can just store information about the platform direction without the cue, then it probably wouldn't fail at the task in the first place. The information about the platform direction seems tethered to information about the cue in ways that they the information states will always or usually covary.

So, the Malfunction Approach runs into trouble explaining the task failures from McGauran and Wortwein's cue rotation probe trials. The trouble is that there don't seem to be any malfunctions in the task relevant systems of those rats. If there are no malfunctions, then the approach cannot identify difference makers to the failures. It yields an empty analysis.

To summarize, consider, again, your failed trip to the Pogonip goldfish pond, and suppose that you are determined to blame the failure on some malfunction. What could you say? There are no obvious sources for malfunctions. You do not remember feeling symptoms of illness or injury before setting off. And nothing unusual directly interfered with your trip. There were no trail closures, frightful scenes, or run-ins with toxic plants. What's more is that you traveled about as far as you usually do, at roughly the same pace. Whatever malfunction you point to must fit with these facts.

This gets you thinking about the information you used to guide your travel. Deep in the redwoods, with the clocktower out of sight, you look at your special beacon-device. As far as you can tell from previous trips, it's supposed to light up whenever the goldfish pond is in the direction of an hour-hand pointing in the direction of three o'clock and, by coincidence, when the clocktower is dead-ahead. You do a quick 360-degree spin just to see if it works at all, and see a quick blip of light. No malfunction there. The devices still functions to carry information. You wonder if, maybe, the malfunction has to do with accuracy or truth. But then dismiss the thought as irrelevant. GPS maps and travel instructions

can be accurate or true, but not a beacon. It's just not that kind of thing. Puzzled, you use your GPS device to check the beacon and learn that, for the entire trip, the beacon had been lighting up when the goldfish pond was in the direction of an hour hand pointing at 6 o'clock. It's not doing, so far as you can tell, what it is supposed to be doing. And you conclude that the beacon is malfunctioning. Satisfied with having found a malfunction to blame the trip on, you text a fellow goldfish pond visitor about the faulty device. They reply "no, no, no, that is not the beacon's function! These beacons have the function to carry information about the clocktower. Don't you see that is when the beacon always lights up? When the clocktower is dead ahead? And look, it still does that perfectly. There is no malfunction. You just didn't realize the city moved the clock-tower!". Despite your efforts and determination to find a malfunction to blame the trip on, you just cannot seem to find one.

Section 3

The Misrepresentation Approach takes the same starting point as the Malfunction Approach: rotating the extramaze cue causes something to happen to the rats' internal, task relevant systems and it fails the task as a result. However, the Malfunction Approach is a more general strategy that promises to explain how any system with a history of design, selection, or learning becomes involved in success and failure. That is why fruitful applications are found across domains

like computer science, psychology, medicine, civil-engineering, cellular biology, and even practical, everyday domains like cooking or woodworking. The Misrepresentation Approach is a little different. It involves saying the cue manipulations cause something to happen to distinctive, *representational systems* inside the rats, which then causes the rats to fail. One consequence of the emphasis on distinctive systems is that it is more specialized, and fruitful applications of the approach seem restricted to psychology.[49]

The distinctive representational systems are characterized by special jobs to read or write special objects that, themselves, have special jobs to tell a story or paint a picture about how things are elsewhere.[50] Those objects might tell a story/paint a picture about what is going on elsewhere in the organism's body, like metabolic representations of fat stores (or fuel) thought to give migrating sea mammals clues about when to head back to feeding or breeding grounds.[51] Or, they might tell a story/paint a picture about what is going on in its environment, like its perceptual representations and corresponding memories.

---

[49] To be clear, I am not saying the Misrepresentation Approach is, in fact, a psychological theory. I am saying that the kinds of systems it references are so distinctive that reference to them seems characteristic of a special science because it uses posits that cannot be reduced to posits of other sciences. See, Fodor 1974, for example and Millikan 1999 for discussion.

[50] My point here is that it is helpful to distinguish between representational systems and representations. We can characterize both functionally. Representational systems have the function to produce/consume representation objects, and representation objects have the function to be about other things.

[51] These types of representations are thought to give the 'head home' cue to pregnant elephant seals, who time their thousand-mile journeys to shore so that they make it back within three days of giving birth. Condit, 2021 and Beltran, 2022.

Application of the Misrepresentation Approach involves saying those systems do an adequate job of reading/writing those objects in training. They produce or consume objects that tell a story/paint a picture of how things really are in the maze, about where the platform really is relative to the rat's location. And that they do a bad job in the cue rotation trial. The very same systems read/write an object that tells a fictional, made-up story about how things are. The object *misrepresents* the relationship between the rat and the platform. It is like when bad directions give you the wrong idea about the location of your destination. The approach says we can trace the failure back to a misrepresentation in the rats' task relevant systems.[52]

The approach encapsulates a familiar idea: there are 'things' (though I will keep calling them "objects") inside intelligent creatures like us that are about other things, and those objects play key roles in predicting, explaining, and understanding each other's behaviors. I predict that my friend will arrive at the airport today at 4pm, and I explain my own behavior of driving to the airport at 3:00pm with an empty passenger seat, in part, by reference to their belief about when they will arrive. That is, I understand my friend's actions, plan my own, and act appropriately, in part, by attributing something inside of their head that tells a

[52] I'm not so interested in analyzing 'aboutness' or intentionality here. I think our ordinary understandings are more than enough for us to do philosophy about the science of mental representations (because that's what those scientific concepts are based on anyway). Some other analyses of aboutness that I really like come from Stebbing, 1926 (p.30-38) and Siegel, 2021 (the newspaper/bucket analogies). And I treat the category of things with aboutness as a 'hurly-burly' category. We can identify sufficient conditions for something having aboutness, but identifying necessary conditions seems too problematic.

story about something else: the time they will arrive at the airport. I do not seem to predict or understand things *just* in terms of the written content of what they say or send. Just think, if I knew my friend lacked a belief they were going to arrive at 4pm, maybe because they are a prankster who sent the message as a trick, then I wouldn't waste my time driving to pick them up. In fact, I would predict, understand, and react to their message differently just because of these more malicious belief and intention objects I think are inside their head.[53]

That is not to say the Misrepresentation Approach is a ringing endorsement of our ordinary practices. Just that it preserves (at least) a small slice by saying there are objects inside the rats that are like beliefs in that they are about how things are, but it does not say things are as easy as looking into the rats' belief systems for false beliefs (at least not as we ordinarily understand them). What's more is that the Misrepresentation Approach says representational objects are characterized by three fairly technical and 'scientifically' characterized properties. It tells us the inner objects (representations) have conceptual vehicles, have conceptual content, and *can* have content that is false.

As an aside, we may wonder how the representational objects we posit inside of us, characterized by these properties, are meant to be related to the things we think of in our ordinary explanatory practices, like beliefs. Ultimately, it depends on *what else* you think about beliefs and the things in our ordinary

[53] Fodor (1987) uses a similar example to illustrate the success of common-sense psychology. My point is a little different: attribution of representation objects is something we are all familiar with (perhaps because of its success).

explanatory practices. If, for example you think some beliefs have original/underived meaning and that the posits of the Misrepresentation could not, then the posits of the Misrepresentation Approach are poor grounds for beliefs (Harnard 1999). But here is one thing that seems true: without getting clear about the nature of beliefs, using them in concert with the posits of the Malfunction Approach seems like a comingling of explanatory strategies or 'ways of looking at things'. It is not unlike using the molecular structure of a table to understand or explain the table's durability. If the property of table-durability reduces to properties of the molecules that make-up the table within the framework of our explanation, then we should expect the comingling of explanations to be fruitful. But if table-durability is not reducible to the properties of the molecules that make up the table, then it is not straightforward how the comingling of explanations will prove helpful.[54] Or, here is another way of making this point: if our explanatory approach limits us to speaking of molecules, it is not so straightforward how we ever get to talking about tables and their properties, like durability. Likewise, if our approach to explaining error limits us to speaking about special systems with special properties, it is not so straightforward how we ever get to talking about beliefs on that approach.

To summarize (so far) the Misrepresentation Approach involves saying the rats' internal, representational systems are (at least in part) responsible for the rats' success and failure at the maze task. The systems read/write representation

[54] Maybe they will, but we would expect some account why.

objects that are about other things, in sort of the same way we think of beliefs as being about other things. However, we should be weary of relying too heavily on the analogy between the Misrepresentation Approach's internal representation objects and the way we ordinarily think about beliefs because the Misrepresentation Approach's internal representation objects are characterized by three properties (it is unclear and beyond the scope of this dissertation whether beliefs are). To further explain what the approach says, I explain each property in turn.

1. Representations have conceptual vehicles.
2. Representations have conceptual content.
3. Representations can have content that is false.

The first property is assigned to representation objects themselves. It is a property of the physical structure that does the representing in a representational system, or, as it is sometimes called, the *representational vehicle*.[55] The strings of letters below can help elucidate what I mean by this.

a. Le_Pto EIdPs

b. Empedocles leaped

c. Empedocles liebt

[55] See, for example, Neander 2017 and Block 2023

Presently, a is not the sort of thing we are interested in. It is a random string of letters, not a representational object. It could easily become one though. All we need to do is introduce a convention like "let's use the string of letters 'Le_Pto EIdPs' to stand for the fact that the cat is on the mat" and, Presto!, the string will represent the cat being on the mat.[56] b and c are already representation objects, and they form a funny pair. They are indistinguishable when uttered, but clearly distinct when written.[57] We can characterize those similarities and differences by saying things like "They both start with the '*Emp'* sound." or "One is a string of 15 letters and the other is 17.". We need not refer to what the representation objects are about. We need not even understand English or German. That is what I want to emphasize. Prima facie, we can characterize properties of representation objects without speaking about their content.[58]

The first property assigned by the Misrepresentation Approach is like the properties we use to characterize the similarities and differences between b and c. It is a property of representation objects themselves, not the stories they tell or the pictures they paint. The Misrepresentation Approach tells us those representation objects are composed of simpler representation objects. But not in

[56] It may be a little more difficult than that, but the point is just that we could easily and *arbitrarily* introduce some conventions to turn this into a representation.
[57] For further discussion of b and c, see Block 2023 & Davidson 1698, who use the strings to draw similar points about contents and vehicles. a is my innovation.
[58] Maybe there is a tight connection between representation objects and representation contents, such that the type of object it is restricts or determines the content it has (our practice of introducing conventions for a would seem to go against that idea). If there is, then the understanding the content seems necessary for characterizing *some* of the properties of representation objects like their format, but it seems irrelevant to things like the number of letters.

a way that's supposed to remind us of the Earth resting on a turtle's back. The decompositional analysis is meant to terminate, eventually, with simple representational object atoms not composed of simpler representational parts.[59] The representational objects are more like complex molecules composed of simpler molecules or atoms. Or sentences, composed of words. They are complex representational objects that are, themselves, composed of other, simpler representation objects.

But that is just part of what it means to say the representational objects have conceptual vehicles. Conceptual vehicles are not composed of just any simpler representational parts, they are made up of *specialized* representational objects with dedicated, discrete functional roles in their representational systems. These objects have a history. They (or their ancestor objects) were selected, designed, or trained to tell certain stories or paint certain pictures (or to tell stories/paint pictures in a specific way). We would not build a conceptual vehicle by re-arranging or refitting the pictures making up a collage or the tiles in a mosaic. The simpler representational object pictures or tiles have no discrete functional roles and can be rearranged in all sorts of ways to create the same or new likenesses.[60]

---

[59] Although, we could always introduce a new convention, like with a, to turn a representational atom into a complex.

[60] We would have to introduce conventions and establish a history of use according to those conventions in order to transform those smaller representation pieces into a conceptual vehicle. We have to 'invent' a new representational system.

What's more is that the complex representations are sensitive to whether the simpler, representational parts carry out their roles successfully or whether they malfunction. Subvert a simpler representation's role, and you jeopardize the complex representation's ability to tell a story or paint a picture in the representational system at all. Not unlike the way subverting the representational function of name phrases like "Plato" or "The Holy Roman Empire" upsets the meaning of the sentences they participate in. If I stop using "Plato" the way it's supposed to be used to refer or uniquely describe and start using it the way we use predicate expressions to describe a sortal like *belongs to Plato*, then that really changes how my utterance of "Plato is mortal" operates in our language system.[61] I no longer tell a story about the person, Plato, belonging to the sortal of mortal things. I tell a confused double-gappy 'story' about two different sortals. And it is completely mysterious how the copula is supposed to relate them.[62] In the same way, changing how the simpler representational parts of a conceptual representation are supposed to represent or 'do their job' upsets the complex representations status as a representation. Conceptual vehicles like these form the

[61] My point is about *how* these vehicles represent, not *what* they represent. It doesn't matter if we subvert the function by having "Plato" refer to the person Aristotle. The representation we have in mind would (likely) be false, but it would still work as a representation. So, its not a matter of subverting content functions, but of subverting more basic representational functions.

[62] I take this point to coincide with Russell's (1905) point that many naming expressions like "Plato" do not, in fact, operate as devices of direct reference and that they actually operate as unique descriptors. Russell's point is that treating them as devices as direct reference is a little like subverting their representational function (it's a misunderstanding of their representational function), and that thinking about them that way upsets their ability to represent certain matters like non-trivial identities, hence why we run into Frege problems. For discussion, see Heck (2006) and Perry (2020).

building blocks of Fodor's classic Language of Thought hypothesis (Fodor 1987, Fodor and Pylyshyn 1988) and modern versions of the hypothesis (Quilty-Dunn 2023).

Mental files are also good examples of representation objects with conceptual vehicles. Here is Jeshion's (2010) analysis of their role in singular thoughts (thoughts about an object of direct perception).

> "Here's a natural way to construe the essential singularity of thought from mental files: Thinking of an individual from a mental file just is thinking of an induvial with a mental name or demonstrative. And, because thought with mental names and demonstratives is ontogenetically rooted in the coupling of them with FINSTs,[63] thinking of an individual with a mental name or demonstrative is essentially singular.
>
> FINSTs can, and typically do, go solo in their non-conceptual referencing in the sense that they do not need an accompanying mental demonstrative partner- a mental "that", "she", or "it"-in order to track a single object. But FINSTs can be accompanied by a mental demonstrative, with the mental demonstrative referring to the object that the FINST refers to Arguably, such use of mental demonstratives is necessary for thought (though not tracking) of individuals. Through their use in communication

[63] F.I.N.S.Ts or fingers of instantiation are mental representations that work as a reference or indexing advice. The representations 'point' at an object and track it as the same object as it moves through time and space. Pylyshyn (1989).

> and associated mental processes, mental demonstratives, construed as a type, come to function as mental stand-ins for FINSTs. They develop so as to function constitutively as abstract singular referring devices by means of which we think singularly about individuals. By virtue of this general constitutive function, mental demonstratives can serve as devices of singular thought even in the absence of any perceptual indexing of the object. So long as the thinker has a means of identifying the object-and in the absence of perceptual indexing, descriptive identification serves-mental demonstratives function cognitively to afford singular thought about individuals." Jeshion (2010) p.135 *might try and find a better quote that involves the object files part. P. 132-139

Jeshion claims that mental files are the representational vehicles for singular thoughts; that the mental files are composed of simpler representations like FINSTS and demonstratives; and Jeshion claims that each of those sub representations has a discrete, specialized role to play that is essential to a complex mental file's role in singular thought. Thought about individuals is impossible (arguably) without them! King (2020) offers a helpful summary of how the simpler representation pieces can combine, recombine within the representational system to account for more and more of our singular thoughts.

> "Jeshion holds that agents have mental files that bind together information the agent takes to be about a single individual. An agent's system of files constitutes her perspective on what objects there are in the world and what properties they have. Because this system of files constitutes the agent's view about how objects are individuated (one for each file), the agent updates, merges, separates, and initiates these files in characteristic ways. When the agent receives new information about an object she has a file for, she updates the file with the new information." King (2020) p.90

Representations with conceptual vehicles (like mental files) are importantly different from some other types of representation objects we are familiar with. They are different from what Pierce (1931-1958) called "icons" or simple representations introduced by convention. Our 'baptism' of the representation object in a is a good example. The string of letters that forms it is not composed of simpler, representational units stitched together like in b or c. To be sure, a *could* be a composed of simpler representations, but we would need to introduce more conventions about those representational parts. And the object would lose its status as an icon.[64]

They are also distinct from representational atoms, the simplest representational parts or units a representation system functions to read/write. Analogies to sentences or molecules are apt. Sentences are different from words

[64] For further discussion of icons, see Ramsey 2007 and Millikan 2012

and molecules from elements in the same way the Misrepresentation Approach's complex representation objects are different from representational atoms. The former are complexes made up of simples. The latter are simples that do not break down to further representational pieces.[65]

They are also importantly different from many of the image representations we are familiar with. However, unlike icons and atoms, the image representations I have in mind do have simpler representational objects as parts. The coffee mug image in Fig. 13 is actually a collage of two, simpler image representations from Fig.14, an image of a cup and an image of a handle.

Fig. 13. Coffee mug image

Fig. 14. Cup image and handle image

[65] Icons are atoms are both simple representation objects but they differ in how they get their representational function. Icons get theirs by convention and atoms by a history of operation in a representational system.

What makes the image representations like the one in Fig. 13 different is that the simpler representation parts are not specialized units like the ones in mental files. They do not have discrete representational roles to represent certain contents or to represent them in certain ways. They are unbound by rules or conventions. I can use the simpler, representational parts of the coffee mug representation to represent bridge supports or even the moon (Fig.15).

Fig. 15. Images of a bridge and a house with crescent moon

These uses and more all seem like fair uses in the game of creating likenesses on a 2D plane. I don't need to introduce new conventions or rules for the Fig.15 images to count as representation objects. And using them that way (or any other way I can think of) doesn't thwart an image's status as an image the way using "Plato" like a predicate term thwarted our attempt to form a sentence.[66]

To summarize this first property, we can say the Misrepresentation Approach involves explaining the rats success and failures in maze tasks by reference to representation object inside of the rats that are, themselves, composed

[66] To be clear, I am not drawing a sharp division between all image representation objects and all of those objects with discrete functional roles. My goal is a little less ambitious. I just want to make clear what I mean by conceptual representations by contrasting them with a few examples.

of representation objects with discrete representational roles. That distinguishes the representation objects involved in the explanatory strategy from other sorts of representation objects like icons and atoms, which have no simpler representational parts and images, which have parts but no distinctive roles.

The second property assigned to the representation objects is a property of its content, or what it is about. The sentence "the snow is white", the expression "The Holy Roman Empire", and Picasso's *Guernica* are all about something else. I said earlier they tell a story or paint a picture. We can start to think about the content as the picture they paint or the story they tell. The content of the *Guernica*, for example, is something like the air of grief and chaos following the bombing of Guernica.

This example sounds strange if we reserve the word "content" for propositions, like the one expressed by both of the sentences "the snow is white" and "la nieve es blanca".[67] The idea being that representation objects, properly construed, are proxies for truth-evaluable *genera* (propositions) that intelligent organisms express or wrap their heads around.[68] But propositions are just some of the things that representations can come to be about. Ordinary experience confirms they can also be about objects, like the Coke cans we see plastered on

[67] For a similar distinction between sentences and the propositions they express, see Ayer 1936 p. 32-42. For examples of philosophers who treat contents as propositions see Gluer 2016 and McDowell 1996. For critical discussion of this use of "content" see Reimer 2020.
[68] There is a fair bit of wiggle room with respect to how we may construe the nature of propositions. They might be mind-dependent senses or guises (Frege 1892, King 2020), arrangements of mind-independent properties and objects (Russell 1912), or sets of possible worlds (Lewis 1986). For discussion see Schellenberg 2020 (Schellenberg's focus is on the general-ness, which is very helpful).

billboards.[69] They can be about property instantiations, like the giant copy "REFRESHING" sprawled across the Coke billboard.[70] They can be about fictional entities, possible worlds, or imaginings like *Girl With a Unicorn* or a sci-fi novella.[71] They can even be about mental particulars like a memory of your last experience sipping Coke. Our practical, cognitive, and even emotional engagement with these representation objects takes us beyond the object itself to a wide variety of other things. And thinking about the content merely in terms of propositions impoverishes our abilities to understand and explain that engagement. To abuse an old metaphor, representation objects are like lighthouses that cast beams outward to something else. Importantly, the beams don't always fall on propositions. They fall on things like objects, property instantiations, mental particulars, and more. I will use the word "content" more generally to mean whatever it is a representation is about.

But, however we construe the contents of representation objects,[72] it makes sense to distinguish between contents that are accessible to a target organism (or group of organisms) and those which are not. Think about someone just starting to learn the sport of soccer. It seems unlikely the contents of the representation objects used by fans, players, and officials to think and talk about offsides violations are accessible to the person just learning the sport. In the sense

[69] For accounts of object representation, see Reimer 2020 and Dickie 2020.
[70] For accounts of property representation or property instantiation representation, see Russell 1912 and Burge 2010.
[71] See, for example, Walton 1993, Meinong 1904, or Lewis 1986
[72] Even if we restricted our use to mean propositions.

that they could not, *presently*, access those contents. They could not presently think about them, speak sincerely about them, or read/write representation objects about them.[73] Offsides violations involves complicated rules, and soccer is a complicated game. Or, better, think about a rat. Surely a rat could not read/write representation objects about offsides violations; such contents are inaccessible to it.[74]

We should, of course, expect a fair amount of variety with respect to what counts as accessible between organisms and across species. No rat could understand offsides rules, but lots of humans can (even if they don't seem to at first). We should also expect variety depending on how we understand the notion of "access" (more about this later). But it still makes sense to distinguish between those contents that are accessible to a target organism and those that are not. Even if we disagree about the details.

Not only that, but the distinction is key to some philosophical hypotheses. Conceptualism about perceptual experience is a good example. Conceptualists claim the contents of all perceptual experiences are accessible to the organisms that have them. This amounts to saying that an organism's ability to read/write representation objects about P is a necessary condition for that organism to perceive P. They must 'grasp' the content already in order to see it. Such contents are meant to be distinct from the ones found in the neural systems responsible for

[73] They could parrot or repeat the words spoken by fans, players, and officials, but that is not the same.
[74] Heck (2008) draws a distinction along similar lines.

visual processing, which produce representation objects with content that *we* (perceiving subjects) are not privy to.

Consider the Ebbinghaus illusion (Fig.16), which uses depth cues on a 2D plane to trick subjects into perceiving the orange circles as different sizes. Now consider that some of the first properties of the circles represented in early vision systems are about the circles' *retinal image size,* which is, roughly, the number and distribution of photoreceptors the circles impinge across a retina. Retinal image size is a function of a stimulus's actual size and distance from the retina, like the way its shadow is a function of its size and distance from a light source. Move it closer and the retinal image size will increase; move it further and, the size will decrease.

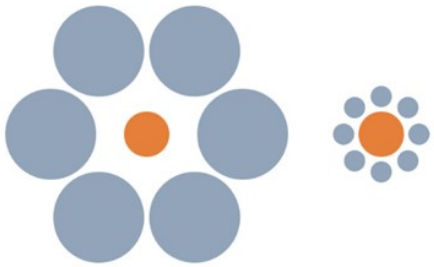


Fig.16 Ebbinghaus Illusion. Thomson and Macpherson, 2017

The story vision scientists tell is that light bounces off of the stimulus and impinges photoreceptors on the retinas. Representations about the size of the retinal perturbations are then carried through channels to the visual cortex where the information is processed and sent out to other systems. And since the circles

are the same size and distance from our retinas, we should expect early vision to represent them to have the same retinal image size. To get a little technical, a representational system in early vision, like the retinotopic map system, would write two representation objects about (perhaps among other things) the photoreceptors perturbed by each circle stimulus. And since the photoreceptor perturbations would be similar in number and distribution, we should expect the content of the representations to be similar too (or similar enough). But we do not see or otherwise *perceive* the circles to have the same retinal image size. We never represent those properties in conscious perception. The content just never makes it to that point. That is part of the conceptualists point. We, as conscious perceiving subjects, lack apprehension or do not grasp the content told by my early visual systems, and we do not perceive them as a result. The messages sent and received by my retinotopic maps are like those sent by my stomach and liver. They are, as Kant would say, nothing to *me*.

> "It must be possible for the 'I think' to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me." Kant, 1781.

The distinction can also be found at the heart of realists' hypothesis in the philosophy of science. Realists about a science object think of those as real, mind-

independent objects that exist independent of our explanatory practices. A realist with respect to atoms, as we understand and model them in physics, thinks those atoms actually exist. As a consequence, they likely hold hope that scientific practice can deliver true representations (models, explanations) about atoms. They think the real nature (or, at least, some part of it) of atoms can be represented with conceptual contents. That its real nature can be represented in ways accessible and grasped by us.

The Misrepresentation Approach tells us that the representations involved in the rats' successes and failures are like the ones mentioned by conceptualists about perception or optimistic realists about atoms. They have contents that are accessible to the relevant organism in question. The contents are not so fine-grained, coded, specialized, or foreign that they count as nothing to the organism.

That does not mean the content is, in fact, accessed by the organism, just that it could be. There are many sentences written in books I could read that I have, in fact, never read. They all express accessible contents I have not yet accessed. In a similar fashion, the rats subsystems may store and use conceptual representations that just aren't or haven't been used by the organism.

Another way of putting this is that the Misrepresentation Approach invokes the kinds of contents we usually find in person level explanations (Dennet, 1996 & Drayson, 2014). It invokes the contents we reserve for things like beliefs, desires, mental models, etc.- things we assign and use to understand persons, not subsystems or parts of persons. To illustrate, consider a fairly

straightforward experiment task asking human research subjects to report whether they saw a Gabor flash across the screen (Fig. 17). What explains a subject's success-using a keystroke to select "yes" when there was, in fact, a Gabor? One story says the subject accumulates evidence about the Gabor, forms a belief, then makes a report based on that belief. Another treats the subject like a system of interconnected subsystems that carry out their tasks by reading/writing coded proprietary messages and terminating, causing the next system in the chain to carry out its task. Like a falling domino striking the next one, the coded representations sent by the retinal image size system could have a direct channel to locomotion (or effect another subsystem with a direct channel to locomotion). A dedicated retinal image size representation reader could terminate in causing the finger-twitch involved in the report. No beliefs, mental models, or other 'important to me' representations and their contents needed!

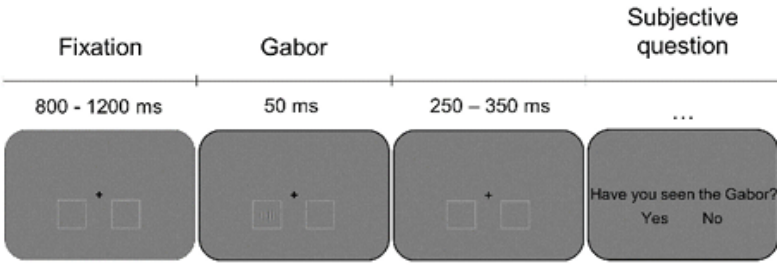


Fig. 17 Near-threshold visual detection task. Participants were asked to report the appearance of the Gabor stimulus. Melcon et al. 2023

To summarize this property, the Misrepresentation Approach involves explaining the rats' successes and failures in maze tasks by reference to representation objects inside the rat. Not only do those representation objects have

conceptual vehicles (meaning they are made up of simpler representation objects with discrete representational functions), but they have a special kind of content. They have content that is accessible to the organism itself. They are not the proprietary, coded contents of representation objects read/written by subsystems. They are more on par with the contents of things like beliefs and perceptions.[75]

The third property of the representation objects identified in the Misrepresentation Approach is also a property of the objects' content. The representation systems identified in the approach read/write objects with contents that have a special property: the contents can be *false*. The notion of 'false' at work here is technical. We can broach it with the more general idea that some representation objects can have content (are about things but not necessarily propositions) that comes apart from how things are. The thing represented or the way it is represented can come apart from how things are. They can exhibit differences or dissimilarities, like in a fictitious tale or an impressionist painting. Contrast these with representation contents that cannot come apart from how

[75] Imagine three people rowing a canoe through a system of streams and channels for the first time. They sit in a row as they row. The first person reads and steers according to Turn-by-Turn directions so that they repeat the directions to themselves and consider them before steering (imagine them mumbling to themself: "hmm, row 100 paces and turn left…do I see anywhere to turn left…I do!"). The second person hears some of the directions, understands, and even considers them, but it is not their job to steer so they do not. Nonetheless, the directions guide where they end up via the person at the front of the canoe. The third person hears none of the directions, but would understand and could consider them if they did. Nonetheless, the directions guide and direct where they end up. The way I am using the word "accessible", the turn-by-turn directions are accessible to everyone in the canoe. With respect to the third person, we can say the directions guided their behavior and that they were accessible even if they did not access the directions. There are no special, coded turn-by-turn directions for navigation and the navigation system.

things are. Logical truths like "Mary Shepherd was Scottish, or she was not" or a guaranteed-to-work live video feed are good examples.

But the idea that the content can slip or come apart from reality is just a rough sense of what it means to say the content can be false. Some contents come apart from how things are but are not false (at least, not obviously). Sherlock Holmes didn't exist, and so a story about Holmes's residence at 221B Baker St. in the 1890's comes apart from how things really were at 221B Baker St. in the 1890's. But the content of the summary statement "Holmes lived at 221B Baker St." is true. Or, at the very least, it can reasonably be understood as true.[76] Fictional contents are not the only example. The content of *Young Woman with Unicorn* comes apart from how things were at the scene of the painting (or how things were anywhere in reality), but it doesn't feel right to call it false (Fig.18). Bad or rounded quantity estimates are another example. My computer may return a poor estimate as False, like when I enter "3.0" instead of 2.987 into a .csv cell. But no one would really think of "3.0" as false when it represents 2.987. Not in the sense we mean when we use the Misrepresentation Approach.

This points us to a distinction between a more general, loose notion for contents that come apart from the world and something like a more specific, technical notion of 'falsity' for contents that come apart in distinctive, special

[76] Here's an argument for thinking it is true: "Holmes" just (or in part) means "the person who lived at 221B Baker St."

ways.[77] This latter notion is what gets used in the Misrepresentation Approach, and the examples above just don't seem to fit that bill.[78]

Fig. 18 *Young Woman with Unicorn*. Raphel 1506

The sharper, technical sense of "false" requires content about satisfaction or *meeting a condition*. The representation object must be about something satisfying something else, like an argument satisfying predicate or an object instantiating a property. If the satisfaction *really* happens (it obtains or occurs out there in the real world), then the representational content is true. If it does not, the

[77] For other distinctions between truth, accuracy, and other kinds of veridicality see Lewis, 1971 and Burge, 2010. For uses of "accuracy", where accuracy is the same as truth, see Siegel, 2010 & 2021

[78] For arguments about other examples see Rescorla (2009) for arguments about robot maps, Camp (2018) for arguments about maps more generally, and Fodor (1987) Siegel (2010) for consideration (not endorsement!) of arguments about perception.

content is false. Well-formed expressions in a language for predicate logic are good examples. $(\forall x)Bx \rightarrow Yx$ is true just in case the satisfaction depicted by the representation obtains. An expression like "All Banana Slugs are yellow" has a similar flavor. When we say it sincerely, we mean something like "all of the stuff in the Banana Slug category satisfies the conditions for belonging to the yellow category". They are also members of that class or group. Beliefs, thoughts, and sentence representation objects can express something similar (so we think). All of these representation objects would be true when, in fact, all Banana Slugs are yellow- when the satisfaction representation by the object is a real actual satisfaction. They are false when satisfaction doesn't obtain.

So, what makes the difference between contents capable of veridicality, more generally and falsity is what they are about. The latter must be about satisfaction. Two points of emphasis. First, falsity is a species of veridicality and so it has to do with the world. Not organism or species fitness. The world is the measure for this property, not the organism and its history. That makes something's being false different from its malfunctioning, even if every misrepresentation is a malfunction. Second, the value False is one of two possible values. Which is appropriate since something either satisfies something or it does not. There is no middle ground.

To put it all together, we can say the Misrepresentation Approach is an explanatory strategy that involves saying there are representational objects inside the target organism. The representational objects are products of specialized

representational systems that function to read/write the objects, and they are characterized by three properties. They have conceptual vehicles, meaning they are complex representation objects made up of simpler representations with functions to represent specific types of content. They have conceptual content, meaning they have content that is accessible to the target organism. And they can have false content, meaning that the satisfaction or instantiation represented by the content need not obtain. The approach promises to satisfy the constraint listed at the top of the chapter by telling us the task-relevant system is a special representational system that participates in success and error. The system participates in success by reading/writing True contents about the location of the platform, which the organism follows to the actual platform location. It participates in failure by reading/writing False contents about the location of the platform which the organism takes to the wrong location.

The approach seems to work for explaining failures by human subjects in *virtual* Morris water maze tasks.[79] Many of these studies involve training human subjects on a virtual counterpart to the standard maze task, then giving them a special *missing-platform* probe trial in which the safety platform is removed from the task.[80] Researchers record the amount of time (in seconds) subjects spend across the maze's four quadrants. Successful performances involve taking a direct

[79] For discussion of virtual Morris water maze studies see Thornberry et. al 2021 and Woolley et. al 2015. For a *non*-virtual study involving human participants in an arena, see Fitting et. al 2007.
[80] Training human subjects involves *n* blocks spread across 1 day instead of the *n* blocks spread across *n* days for rodents. Interestingly, rodents do much better than humans at learning the task. See Schoenfeld et al 2017 for comparison and discussion of humans and rodents.

route to or spending the majority of trial time in the 'correct' quadrant that contained the platform during training. Failures involve taking a B-line to the wrong quadrant or spending too much time in the wrong quadrants.

The Misrepresentation Approach tells us the failures are due, in part, to specialized representation systems in the human research subjects. Those systems (or maybe just one of them) read/write complex representation objects composed of simpler representation objects with discrete jobs to represent specific features of the maze. We could reasonably think of the objects as beliefs, thoughts, memories, mental files, or mental models. Whatever the details, the representation objects carry a message about the platforms location that the human subject can, themselves, understand and comprehend (even if they do not, in fact, presently understand and comprehend it). The subject (or a relevant subsystem of the subject) consults the content of the representation to guide their behavior. During training, the content was true. The believed, remembered, or otherwise thought about platform location was satisfied by the actual location of the virtual platform. But, in the missing platform probe trials, the representation is false. The represented platform location comes apart from the actual location of the virtual platform. The subject isn't privy to this and so follows the content of their representation to the wrong maze quadrant.

Can a similar story be given for the rats? I don't think one can. There are challenges to treating the task-relevant systems as representation systems that read/write representation objects with these technical and scientifically

characterized properties. I identify challenges for each property the Misrepresentation Approach assigns to those representation objects.

The first has to do with the properties assigned to the vehicles of information-the things that are supposed to do the representing. In Morris water maze studies, the relevant vehicles are usually patterns of neural activity. Researchers are interested in studying repeated fluctuations in electrical currents and treat them as carrying information relevant to the maze. Researchers interested in the vehicles of place cell representations are interested in the patterned activity or firing of place cells. The neurons themselves are not the vehicles. They are conduits for the vehicles, making them more like speakers or instruments than a buoy or traffic cone.

The Misrepresentation Approach tells us that representational vehicles decompose into simpler, representational parts, but there is a challenge to thinking about some of the relevant kinds of neural activity this way. It does make sense to say the activity of an array of neurons or a brain structure decomposes into smaller, information carrying parts. We can think of the activity of an array of boundary cells as the combined activity of each boundary cell. The strong signal detected would be like the roar of a crowd, with each individual neuron contributing weaker signals. We can also think of the pulsing signals detected across the hippocampus as the result of activities of the place cells, boundary cells, and head direction cells that make up the hippocampus.

However, much of the science focuses on the information contributions of single-cell neurons like place cells, boundary cells, or head direction cells. The neurons can surely be decomposed into cell parts with activity, but the decomposition doesn't seem relevant to information. A boundary cell that fires when the rat is near a maze wall does not fire because the activity of the cell parts represent parts of the wall. A similar point applies to the length and strength of neuron activity. We can break down and math patterned cell activity like the complex spike pattern below of a place cell (Fig. 19). We can point to the start of the spike, half the strength (measured in microvolts μV), and even isolate a few spikes, but those would only ever amount to artifacts of the signal. We would not count them as simpler information parts because they do not, on their own, seem to be about anything. Just think, a researcher who detects a blip here or there would treat them as noise, not concepts. Researchers are interested in patterns of activity that deviate from a cells usual pattern of activity, and the information buck seems to stop there.

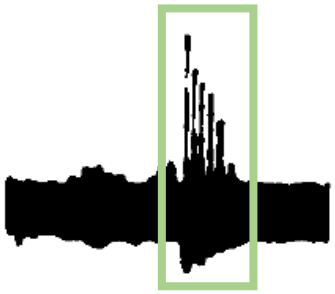


Fig. 19. Signal detected from a place cell neuron across 5 sec. in the CA1 region of the hippocampus of a rat. The place cells default setting is illustrated on the left-hand side of the green box and its complex spike activation pattern is illustrated inside the green box. Ranck (1973). P.472

So the first challenge for using the Misrepresentation Approach to explain failures is that its focus on conceptual vehicles limits what it can identify as difference makers. It cannot identify the activity of single neuron cells as representations or misrepresentations because they are not the right kind of representation objects with simpler, representational parts. The approach can only identify the downstream, coordinated activity of arrays or signals as misrepresentations. That is a real challenge since much of the science focuses on the contributions of individual cells. It seems weird to say things never go wrong or that failure never starts at that level.

But maybe that is an okay bullet to bite in exchange for using the Misrepresentation Approach. Application can be focused on the activity of arrays of neurons or more complicated brain structures like the hippocampus. Except there is another problem with thinking about these as conceptual vehicles. The simpler, representational parts of conceptual vehicles are supposed to have discrete functional roles. They are supposed to carry specific kinds of content like how "Plato" has the function to refer to/uniquely describe the person Plato. The activity of neurons is not so specialized. The hippocampus provides us with good examples. The activity of place cells, head direction cells, and arrays of them will sometimes terminate their functions mid study and pick up new ones (Ludvig 1999). A place cell will stop functioning as a place cell and start inhibiting or exciting other neurons. Not only that, but they are sensitive to non-spatial cues like odors, novel objects, and enclosure shape. In a review of this phenomenon,

O'Keefe and Krupic (2021) categorize over 50 studies that document place cell neurons in the hippocampus changing their firing behavior to respond to non-spatial cues like odors or novel objects. They summarize their review:[81]

> "All theories agree that hippocampal pyramidal cells represent both spatial and nonspatial information. One important difference between cognitive map theory and the others lies in their predictions about the relationship between these two classes of information. Cognitive map theory states that the concept of "place" is generated in the hippocampus whereas nonspatial inputs are generated elsewhere and projected to the hippocampus, where they are embedded in place representations. Pure nonspatial responses independent of the animal's locations may exist in the hippocampal formation as inputs from elsewhere, but even this remains unproven because, in general, experiments that have reported these have failed to test their dependence on the animal's location. That is, the "pure nonspatial responses" are actually covert feature-in-place responses."
> O'Keefe and Krupic, 2021. P.1428

So, here are the challenges associated with thinking about the difference makers as representational objects with conceptual vehicles. Conceptual vehicles are

[81] O'Keefe and Krupic (2021) argue that non-spatial information is important information for navigation and that this shouldn't count as evidence against the hypothesis that the hippocampus functions to construct a cognitive map.

supposed to decompose into simpler representation objects with discrete functional roles. But that doesn't fit with the way the science talks about the difference makers. Some of them lack simpler representational parts, and the ones that do have them lack parts with discrete functional roles.

But those are not the only challenges for the Misrepresentation Approach. There is a challenge to thinking about the representation object as having conceptual content-content the rat itself could read/write. The challenge has to do with the very notion of 'conceptual content'. In short, it is not a good concept (I will start using the word "notion" again to avoid confusion) because it doesn't have anything resembling good sufficient conditions. It does not admit of a working definition for researchers. That presents a challenge. How do researchers know when to apply it to a difference maker and what evidence could they possible provide?

For humans, there are good, reliable markers for when the content of a representation is accessible to them.[82] When they can talk or otherwise communicate about it, the content seems accessible. It seems to us like the person, and not a subsystem of the person, can read/write representations with that content. When, for example, is the content of the offsides rule accessible to a human? It is never a sure bet, but abilities to explain it or play strategically with in

[82] Sometimes the word "accessible" is used technically. See, for example Siegel 2010 or Block 1995. I mean it more generally, in a non-technical way. That's part of my point, actually. It makes sense to use this in a technical way for humans, but it doesn't seem to be usable in a technical way for animals like lab rats.

it are good evidence (prioritizing it vs. other rules, using it to catch other players offsides are evidence of inferences from understanding of the rule in the broader context of the game).

Maybe some horses or dolphins could give us this sort of evidence with respect to mathematical sums or their desires for food, but it would not help us think about the lab rats used across these experiments. The evidence is too tied to linguistic behavior like issuing written/spoken statements or drawing inferences. Lab rats just don't exhibit that kind of evidence. Not during studies and not during their free time in their storage pens.[83]

Where else could researchers turn for this evidence of access? They could look at the rats' brains. Here are two strategies worth considering. The first involves looking at brain architecture and the second involves looking at the activity associated with the representational content. Let's start by thinking about architecture. Generally, claims about architecture are claims about where things are and how they function relative to other parts of the brain. The idea here is that there may be an area of the brain that is responsible for access. There could be something like a belief box or a desire box (Schiffer 1981), and content would be accessible to the rat if it could be 'placed' in the appropriate kind of box. It could even be determined by neuroscientists where the belief box areas are and whether a representation with the appropriate content is in it.

[83] Chater and Heyes 1994 regard this as evidence that conceptual content cannot be assigned to animals. They write "no clear sense has been provided for the claim that nonlinguistic animals have concepts'.

Another idea is to look at the neural activity associated with the representational content. Typically, the more neurons that activate in response to a stimulus, the stronger the signal associated with the stimulus. Stronger signals are prioritized by organisms on this picture and, once signals get strong enough to meet a threshold, they are deemed 'important' and broadcast to the organism. This idea is central to Signal Detection Theory, which uses statistical representations of behavior to determine the threshold for organism level detection (Fig. 19).

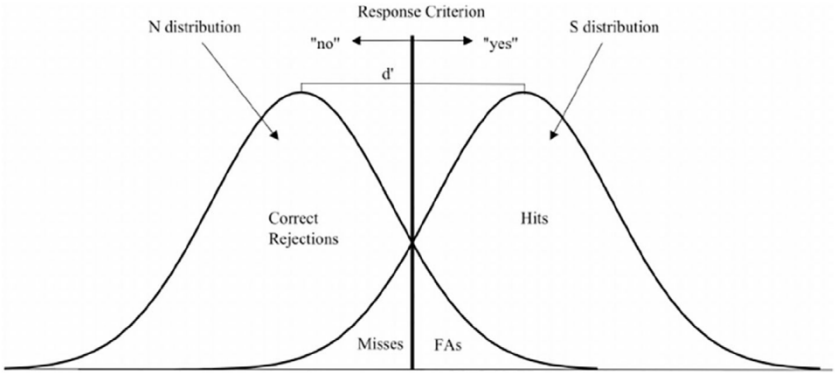


Fig. 20. Model for determining the response criterion on Signal Detection Theory. MacMillan and Creelman 2005.

Both of these strategies tell us to look to the brain for evidence of access. We can find both at play in Global Workspace Dynamics, the neural counterpart to Global Workspace Theory. Global Workspace Dynamics tells us that the cortex and thalamus are sort of like belief boxes. They function to strengthen signals collected by subsystems and broadcast them constituting a 'neural workspace' so

the signals are accessible for conscious perception and thought (Barrs 2021, DeHeane 1998).

The problem with relying on the brain is that the relevant architecture or activity only seems involved with access because they are associated with the linguistic behaviors that reliably indicate access. Access is not a neural notion. It has to do with an organism's abilities and organisms are more than their brains. Claims about architecture and strength of signals can indicate conceptual content, but only because they have been corroborated by the linguistic behaviors. In lab rats, the relevant architecture or signal claims could never be associated with linguistic behavior. Researchers would never see anything like a report or clear, demonstrable evidence of rule-following or inferences. They would just report a content in the belief box or broadcast during the task and that's it. There would be no connection to access.

To illustrate, think about the concept of 'tasty'. It is not a very clear concept, but we get along well enough with it. Partly because we associate it with behaviors like indulging, eating more, or paying more for something. Now suppose a chemist is interested in tasty foods and discovers that most of them have high sugar content. How do they discover that? By discovering connections between the foods that elicit the tasty behaviors and their high sugar content. If the tasty behaviors drop out of the pictures, there can be no connection between a food's being tasty and having high sugar content. If aliens landed and tried our

high sugar content food, we could not promise they would be tasty. The high sugar content, on its own, is not a reliable indicator of tasty.

Here is another way to frame this point. The notion of conceptual content makes intuitive sense, but we run into trouble applying it. Luckily, in the case of humans, we have something close to reliable sufficient conditions for the notion: linguistic behavior reflecting the content. Thanks to recent work in neuroscience, we also know about the neural processes likely involved in that behavior. Indeed, those neural processes are so involved that they become indicators the sufficient condition for conceptual content is met. But, with lab rats, the notion of conceptual content is even more challenging. The reliable sufficient conditions for conceptual content aren't there. And so the architecture or signal claims are just that: claims about architecture or signals. They don't bear on conceptual content at all.

The challenge for thinking the information objects have conceptual content comes in finding reliable sufficient conditions for when information is conceptual or not. It is a problem with the concept or notion of conceptual content. That is distinct from the problem for conceptual vehicles, which involves a clear concept or notion but has an empty extension when applied to Morris water maze studies.

The challenge for thinking the information objects have content that is capable of being true/false is abductive. Thinking about the representation objects another way is more explanatory, so why not think about them that way instead.

The Misrepresentation Approach tells us to think about the relationship between information and the world in terms of satisfaction. The information content stipulates conditions that are either satisfied by the world and true or that come apart from the world and are false. A content either matches the world or it doesn't. There is no in-between or middle value.

But what if we thought about content-world relationships a little differently. We could think about contents as being more or less accurate of the world. Contents don't just come apart from the world, they diverge in degrees. An artist's drawing of my dog and my own will not, strictly speaking, satisfy the conditions for being true of my dog, but one will be more accurate. Thinking about the information objects as being able to come apart from the world in degrees and using the terms "accuracy"/"inaccuracy" to capture this creates an explanatory advantage. It does a better job of explaining task failures.

Not all failures are equal. The rats fail cue rotation trials and other kinds of probe trials in different, predictable ways. In cue rotation trials, they follow the rotated cues. Rotate the cues 30 degrees to the left and the rats veer to the left by about 30 degrees. Rotate the cues 90 degrees to the right and they veer to the right by about 90 degrees. One of these failures is worse than the other. Rats who veer off path by 5, 10, 30 degrees are closer to the platform and spend less time searching for it and escaping the maze. In terms of objective measures like swim latency or swim distance, they do not fail as bad as rats who veer off path by 90 degrees.

We could just say the rats' information is false, that it comes apart from the world and that is why the rats fail. But it is better to say the information can be more or less accurate of the world and that the conditions stipulated by the content can either come close to being satisfied or vary greatly from the world. When the rats' information is a little inaccurate, the failure isn't as bad. Minor inaccuracies may even produce success. But, when it is very inaccurate, it leads to a bigger screw up. Speaking about the information in terms of accuracy/inaccuracy captures this connection between degrees of veridicality and degrees of failure.

So, there are three challenges to using the misrepresentation approach to help a hypothesis for Morris water maze success explain the sort of systematic task failures reported in the cue rotation studies. The challenges are conceptual in nature and are meant to challenge the way we interpret and explain misrepresentation in Morris water maze studies. The first set of challenges has to do with the Misrepresentation Approaches' focus on conceptual vehicles. Conceptual vehicles are supposed to decompose into simpler representation objects with discrete functional roles. But that doesn't fit with the way the science talks about the difference makers. Some of them lack simpler representational parts, and the ones that do have them lack parts with discrete functional roles. The second has to do with the approaches' focus on conceptual content. Conceptual content is not a good notion for Morris water maze studies. There are no good sufficient conditions for categorizing conceptual content. The challenge is to come up with those conditions. The third challenge is abductive. Thinking about

the representation objects as capable of accuracy/inaccuracy is more explanatory than thinking about them as capable of being true/false, and so researchers should think about them that way instead.

Consider your failed trip to the Pogonip goldfish pond again, but imagine that your dog always accompanies you. After learning the city moved the clocktower, you think to yourself "I had a false belief about the location of the pond, relative to the clocktower". That gets you wondering. Did your dog have anything like a false belief? It trotted in front of you the entire way, maybe it misrepresented the location of the goldfish pond too.

You start by thinking about what's going on inside your dog's head. Maybe it has things like thoughts and memories. But you are interested in thinking about this scientifically, so you conclude that lots of neural activity is going on in your dog's head. But neural activity, so far as you think about it scientifically, doesn't fit with your way of thinking about representations. Representations are made up of concepts. They are made up of simpler representational vehicles with discrete functional roles. The neural activity in your dog just doesn't fit the bill. The activity of your dog's single cell neurons are like information atoms without parts and the activity of all of your dog's neurons seem to repurpose and change depending on context. There are no discrete functional roles. Not only that, but representations are supposed to be conceptual. Your dog must be able to read/write (token) representations with that content. Is the information your dog uses conceptual? It seems impossible to say. Your dog

cannot exhibit the kinds of linguistic behaviors we take as evidence of conceptual information. It seems consistent with all of this to say your dog's information could be false, that it could fail to satisfy the way things are. However, it seems even better to say that it was more or less accurate of the location of the goldfish pond. The less accurate it is, the further off-track it would get (usually) and the farther it would end up from the pond. It seems the way you think about representation and misrepresentation presents challenges to assigning them to your dog.

Conclusion

An explanation of success at the Morris water maze task must satisfy the constraint from the top of the chapter. It must also identify difference makers to the systematic failures reported in cue rotation studies like McGauran and Wortwein's. As I hope to have argued, this turns out to be a philosophical challenge that puts pressure on many of the ways researchers talk about and conceive of failures and error. I problematized three, standard approaches to treating the relevant navigation systems so that they can participate in both success and failure. Treating the systems as though they hold fast between success and error, with all blame falling in the environment, does not work. Researchers must conceive of changes in the content, functions, or veridicality of the systems. Treating them as though they malfunction does not work either, since there are no discernable malfunctions in the systematic task failures. Finally, treating them as

though they misrepresent does not work because the systems do not seem representational in the first place.

In Chapter 4, I look at ways of talking about non-conceptual representational content carried by non-conceptual vehicles. The proposal comes in the context of thinking about differences between nearby and faraway visual cues. In future work, it will be applied to explaining the task failures discussed in this chapter.

Chapter 4: Rethinking the Concepts of "Distal" and "Proximal".

Researchers often explain rats' performance in the Morris water maze by reference to visual cues. What's more is they claim rats use distal (far-away) visual cues differently than they use proximal (nearby) ones. For example, Hébert et. al. (2017) report that removing distal cues completely disrupted rats' ability to complete maze tasks, while removing proximal cues had no effect. Similar claims are widespread throughout maze studies, despite variations to experiment task, protocol, and even the species of animal.[84]

Of course, drawing conclusions, making predictions, and designing/using experiments about distal and proximal cues involves a common understanding of what they are. In this chapter, I characterize the dominant working definitions, where the cues are defined by their location relative to the boundaries of the Morris water maze. Proximal cues are inside the maze, like stray marks on the maze's interior walls. And distal cues are outside the maze, like posters on the laboratory walls or lightbulbs hanging from the ceiling. Then, I argue the working definitions do not allow for claims about distal and proximal cues to generalize to real-life navigation behaviors, like long-distance migration. The maze boundaries do not and cannot exist in the natural environments where these behaviors take place.[85] So, there can be no proximal cues within maze boundaries or distal ones

[84] See, for example, Vorhees and Williams 2006 and Craig et al. 2005. More examples will be explained in detail in section 1.
[85] I will argue that, by definition, these devices distort an environment by reducing confounding variables and introducing observation. P.6

beyond them. Such cues just cannot exist in those environments, and so they cannot be difference makers to behaviors in those environments. It follows that claims about them cannot play any explanatory or predictive role. They are not claims about difference makers. In light of this, I argue that researchers should look for another set of definitions that work for claims about natural environments. I consider two proposals.

First, I consider definitions that draw the boundary between distal and proximal cues based on information carried by boundary cells found in the hippocampus and subiculum of rats (Knierim and Hamilton 2011 & Lever et al 2009). This approach has curb appeal because it is intuitive and promises to avoid the problem I present for the dominant working definitions. However, it faces other problems that have to do with the specifics of boundary cell function, lack of boundaries in sparse environments like deserts, and long-distance navigation behaviors like scavenging that take a rat away from its learned boundaries. The proposal is informative, but the problems should move researchers to try something else.

The second proposal builds on the first in that it uses a rat's information to mark the difference between distal and proximal cues. However, it takes information about a cue's retinal image size to be the relevant kind of information. The idea is that information in early visual processes about a distal cue's retinal image size remains the same (or relatively similar) as a rat moves around its environment. Those cues are too far away for visual systems to detect

differences in the image size (Ingram et al 2016 and Soma et al 2012) or are modulated by top-down processing mechanisms (Zeng et al 2020) as the rat moves around. Information about a proximal cue's retinal image size, on the other hand, changes as the rat moves around its environment. These cues are close enough that the visual system detects changes in its image size as it approaches/moves away or that modulation effects aren't as severe. I argue that, on this way of thinking, claims about distal or proximal cues make a difference to explanations of natural behaviors while also preserving the way the terms are cashed out in laboratory settings. In short, the definitions help avoid a problem while also fitting with current science about Morris water maze studies.

The chapter is divided into four sections and a conclusion. In section 1, I describe the dominant working definitions of "distal" and "proximal" and separate them from other, less popular working definitions. In section 2, I present my argument that the working definitions do not allow for claims about distal or proximal cues to generalize to real-life navigation behavior in natural environments. Following this, I argue that researchers should seek another set of definitions for natural environments. Sections 3 and 4 contain my proposals for new definitions. In Section 3, I consider whether information carried by boundary cells can help us think about the difference between distal and proximal visual cues and argue that it raises too many problems. In Section 4, I recommend that we think of the distinction in terms of the size information carried by the organism's visual systems. I argue that this way of thinking promises to capture

the distinction in a way that is explanatory while also fitting with current science. I end the paper with a brief conclusion.

Section 1: Working Definitions of "Proximal" and "Distal".

Here are the working definitions Knierim and Hamilton (2011) identify and use in their review of research on the different roles that proximal and distal cues play in hypotheses about behavioral tasks and the neural correlates of those tasks.

> "For the purposes of this review, we shall operationally define distal and proximal cues according to the common working definitions in laboratory experiments; that is, distal cues are the cues on the walls of the lab or otherwise removed from the behavioral apparatus, whereas proximal cues are those cues that are part of the apparatus itself." Knierim and Hamilton, 2011 p.1246.

The same working definition is given by Young et al. (2006) for distal cues in Morris water maze experiments.

> "Ideally, the location of the hidden platform is learnt through its relationship to distal cues which are located outside of the pool environment, possibly through the development of a cognitive spatial map." Young et al, 2006

To summarize, proximal cues are located inside an experimental apparatus, like a food reward in a radial arm maze or marks on the inside of a Morris water maze wall (Fig.1). Distal cues, on the other hand, lie beyond the experimental apparatus. They're features of the room the experimental apparatus is housed in or they're contrived features set up outside the boundaries of the experimental apparatus.

Examples of these working definitions are found throughout research on animal navigation and spatial learning. In a guide to basic Morris water maze task procedures, Vorhees and Williams (2006) refer to objects outside of the Morris water maze walls, like paper cutouts of shapes, as distal cues and features of the maze, like welded seams in the plastic tank walls, as proximal cues (Fig.20). In an experiment that measured the effects of conflicting rotations of proximal and distal cues on rodents, Yoganarasimha et al. (2006) refer to posterboard and Styrofoam cylinders placed on the laboratory floor outside of a circular track experimental apparatus as distal cues, and they refer to the different surfaces of the track apparatus as proximal cues (Fig.21). Examples date as far back as O'Keefe and Nadel's (1978) critical discussion of Hebb's (1938) claims about the roles distal cues play in navigation.[86]

[86] O'Keefe and Nadel offer a reason for that operationalization in the last sentence of the passage. I expand on this in section 4.

> "Hebb, for instance, felt that information from the distal environment, rather than the test apparatus itself, was crucially important in defining places and allowing for general orientation […] In our view distal cues are important in specifying directions […] distal cues, by themselves, cannot distinguish amongst places in that environment. Places would seem to be defined by extra-maze cues which are close enough to the animal." O'Keefe and Nadel, 1978 p.72-73.

Researchers so frequently use these definitions that the terms "distal" and "proximal" have become synonymous with terms like "extramaze cues" or "cues inside the maze" (McGauran et al. 2004 and Sanchez et al. 2016). A good example of this can be found in Sullivan's (2010) description of the basic Morris water maze task and protocol,

> "The water maze is an uncontrolled open field maze that consists of a large circular pool filled with opaque water. It is placed in a room containing a discrete set of fixed distal (i.e., external to the pool) visual cues. When placed into the pool, a rat will attempt escape, and thus swim about the pool." Sullivan, 2010

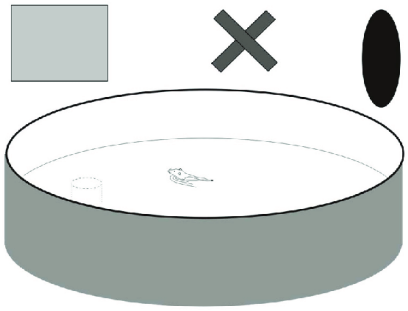


Fig.21. Diagram illustrating placement of distal cues in a Morris water maze. The shapes outside the maze pool are distal cues, while stimuli inside the maze pool are proximal. Save and Poucet (2000).

Fig.22. Photograph showing the circular track with used in Yoganarasimha et al. (2006)'s study. The posterboard and Styrofoam cylinder outside the track are distal cues, and the different patterns on the track are proximal

To be sure, there are some other ways of thinking about this distinction. Carman and Mactutus (2002) use the term "proximal" for hidden cues, like the submerged safety platform in a Morris water maze, and they use "distal" for visible cues, like the walls of a maze or the walls of the laboratory a maze is

housed in. Rodrigo (2002) uses the terms to pick out cues that are near (proximal) or far (distal) from an organism's goal so that the difference depends on where the organism's goal is in its environment. These ways of thinking about the distinction are not very popular, and it's informative to understand why.

On Carman and Mactutus's way of thinking about the distinction, proximal cues and distal cues are not used differently by organisms because hidden, distal cues are not used at all. On Rodrigo's working definitions, distal and proximal cues are defined by their relation to an organism's goal rather than the organism. Cues that are far from an organism but near its goal would be proximal and cues that are far from the goal but near the organism are distal. That is a complete inversion of the way researchers typically use the concepts of distal (far) and proximal (near). Not only that, but it depends on some *other* way of distinguishing between cues that are near an organism's goal and cues that are far from an organism's goal. Without any criterion for distinguishing between near and far, Rodrigo's definitions would not make a consistent and reliable difference to explanations since researchers will be left to color in their understanding of that difference.

To its credit, the standard way of defining the cues identifies a difference that seems to reflect real differences in how those cues are used by rats. When researchers manipulate distal cues after training and leave proximal cues alone, it affects behavior in interesting ways. Rats display a travel bias toward the rotated distal cues (McGauran et al. 2004 and Craig, et al., 2005), even under stress

(Warner et. al., 2013) or after lesions to navigation centers like the hippocampus (Ramos et al., 1998 and Wortwein, 1995). This is taken as evidence that there's a favorite kind of visual landmark that overshadows other kinds. Organisms prefer to use distal visual cues over proximal cues. In addition, manipulating distal cues only seems to affect the direction a rat travels. For instance, McGauran et al. 2004 and Wörtwein et al.1995 report that rats travel the same distance in distal cue-rotation trials, they just travel in the wrong direction. It's like if I swim the same distance I did yesterday, but a drifting buoy led me to swim in the wrong direction. Rotation of proximal cues, on the other hand, doesn't affect the direction of travel (Moses et al. 2002 and Herbert et al. 2017); rats stay on track in these trials. This is taken as evidence that rats use distal cues for information about direction, and that they do not use proximal cues this way. So, in study after study, rats seem to use distal cues and proximal cues differently. They use distal cues preferentially and for information about direction. They use proximal cues if there are no good distal cues or for distance information. In addition, these working definitions provide a straightforward and clear criterion for distinguishing between cues. In most cases, it is relatively clear to researchers whether a cue is outside or inside of a maze apparatus.

To wrap up this section, I will say a little about the difference between this distinction and a distinction from vision science and the philosophy of perception

between proximal and distal *properties*. They use similar terms and get used in nearby disciplines, but are, importantly, different distinctions.[87]

The distinction between distal and proximal properties is used to separate different properties of perceived cues. Distal properties remain constant despite changes to viewing conditions, like the way a coffee mug appears to me to have a constant mug shape and yellow color as I move about and take my seat at a coffee shop. Proximal properties, on the other hand, change as viewing conditions of a cue change. As I move the mug around a coffee shop, there are a huge number of changes to the way the mug surface reflects light from the atmosphere to my eyes, and so the changing reflectance properties of the mug surface are proximal properties.

Vision scientists and philosophers appeal to this distinction for several reasons, like in determining a hallmark feature of psychological explanations (Burge, 2010) or in accounts of the kinds of properties certain cognitive systems are sensitive to (Palmer, 1999).[88] The point I want to emphasize is that this distinction is different from the distinction between proximal and distal *cues*. The distinction between proximal and distal *properties* from vision science and philosophy of perception is a distinction between different properties of cues. A perceived cue like a paper rectangle has both. It has properties like shape, color,

[87] O'Keefe and Nadel (1978) discuss the distinction between proximal and distal in terms of properties in Ch.1 and, without any explanation of the different use of terms, switch to discussing proximal and distal in terms of cues in Ch.3.).
[88] For further accounts of this distinction and its explanatory roles see Orlandi (2014) and Neander (2017).

and quantity that remain constant as viewing conditions change, and it has properties that change like the ways that light is reflected off its surface to our eyes as viewing conditions change. The distinction between proximal and distal *cues* from navigation and spatial learning paradigms is a distinction between cues themselves, not properties of cues. A paper rectangle is either proximal because it is inside of a maze, or it is distal because it is outside of it. It cannot be both.

Section 2: **A Problem for the Working Definitions of Distal and Proximal**

Here is a problem for the working definitions from Section 1: there are no mazes, laboratory walls, or other experimental apparatuses in natural environments. In fact, it doesn't even seem like these things *could* be in natural environments. They are designed by humans to manipulate an organism so that researchers can observe and measure its performances in spatial tasks. Placing apparatuses in a natural environment would distort the environment so that it is no longer natural.

But if there can be no experimental apparatuses in natural environments, there can be no cues inside or outside of an experimental apparatus, just like there can be no historical events pre/post-Santa's birth. It follows that nothing can fit the working definitions and count as distal or proximal cues in those environments. The extension of the concepts is doomed to be empty. We should conclude that, on these working definitions, claims about proximal cues and distal cues cannot be explanatory or predictive of behaviors in natural environments.

Because for something to be, even in part, explanatory or predictive of behavior in these paradigms, it must make a difference to that behavior. A stricter way of saying this is that it must be part of the causal chain that led to the behavior. If I determine that drinking coffee is what makes the difference between getting the jitters and not getting the jitters, then I can say the coffee, at least in part, explains the jitters. If I had the jitters regardless of whether I drink it or not, then I should conclude it doesn't explain the jitters because it doesn't make a difference. If these cues cannot exist in natural environments, they cannot make a difference to behaviors in those environments, and so claims about them are not, in principle, explanatory of any of those behaviors.

This problem illustrates the need for another way of thinking about distal and proximal cues that works for rats in their natural environments. But, before exploring proposals in Sections 3 and 4, it is worth considering a few approaches researchers could take with respect to proposals for new sets of definitions. Imagine a Morris water maze researcher used to the working definitions from Section 1opens a study on elephant seals' abilities for long-distance migration and finds a totally different set of definitions. The definitions make no appeal to experimental apparatus boundaries and seem to help explain and predict the seals' migrations across open oceans. How should they conceive of the relationship between the working definitions they've been using and the newfound ones (especially if they recognize the problem above)? And how they should proceed with their research as a result?

One idea is they should just replace the old working definitions with the newfound ones. The idea being that the old ones cannot explain real-life navigation anyway, better just to replace them. Researchers who take this option would use the terms precisely the same way in laboratories and in nature, whether they are explaining a rat's swim path in a Morris water maze or the migration abilities of seals.

But replacement doesn't seem like the right approach. The working definitions from Section 1 are engrained in the long history of research on Morris water mazes and spatial learning.[89] And with good reason. They are used to effectively explain and predict behaviors and neural activity reported in thousands of studies. What's more, they are incredibly practical. They draw a clear difference between the cues that is easy to measure and communicate about. So, the best paths forward seem to involve (at least) two, distinct sets of definitions for the terms "distal" and "proximal", the working definitions that only work for labs and another set for natural environments.

Researchers could take the relationship to be merely surface level and think of the sets of definitions as unrelated, besides being attached to the same words "distal" and "proximal". When they say things like "elephant seals use distal cues to navigate open oceans", they don't mean to say anything about the boundaries of a Morris water maze or some other experimental apparatus. They just happen to use the same words. It's like when an economist uses the word

[89] For seminal studies, see Hebb (1938) and Tolman (1948).

"value" synonymously with "price", and an ethicist reserves it for things without price. They use the same words but have different and unrelated definitions in mind.

To be sure, there is a plausible case for this. The organisms used in Morris water maze experiments are not natural organisms, in the sense that they are not organisms plucked from nature and placed in a lab. In most cases, their recent ancestors were not either. Most laboratory rats are obtained from animal model facilities like Taconic Biosciences or Charles River Laboratories. They tend to be genetically similar (unless the experiment calls for genetic diversity), and they come from genetic strains like the Sprague-Dawley or Fischer 344 strains of rats, which exhibit desirable traits like being easy to handle (Sprague-Dawley) or susceptible to drug addiction (Fischer 344).[90] These practices, in conjunction with the highly contrived nature of the storage pens and experimental apparatuses used in behavioral experiments might motivate researchers to just say "maybe laboratory rats really do develop different cognitive capacities, and so it makes sense that our definitions are specific to those kinds of tasks.". The idea here is that the abilities and capacities of laboratory rats are different enough from those of natural rats and we may need to use different conceptual tools to explain their behaviors.

I think this is a plausible way to proceed, but it involves revising the way most researchers must think about the evidence collected in laboratories. They

[90] For moral objections to practices like these, see Kitcher (2015) and Singer (1977).

think of studies using Morris water mazes and T-mazes as helping us understand how rats (and other mammals like apes and humans) travel in their natural environments.[91] When researchers discover that rats in Morris water mazes use distal cues for direction information, that is supposed to help their understanding of how other rodents, apes, seals, and even humans navigate in real-life scenarios. So, there are good reasons to explore a closer relationship between the definitions, where claims about how a rat uses distal and proximal cues to navigate a maze task fits with claims about how they use them to navigate the forest floor.

But research on real-life navigation doesn't zero in on the roles of distal and proximal cues the way laboratory research does. There are surely appeals to their explanatory roles.[92] Lots of them. The research just does not involve attempts to control the environment and isolate variables like laboratory research does, and, as a consequence, there are no clearly laid out and widely applied sets of definitions. In the next two sections, I build on theories about rats' abilities to detect spatial differences to propose and consider definitions that promise to work in natural environments and fit with the way researchers use the terms in labs.

[91] See, for example, Frost and Mouritsen, 2006 and Poulter et al. 2018

[92] Some examples include Matsumura et al 2011who explains the role far away coasts plays in providing direction information to migrating elephant seals and Vincze et al 2015 who explains the role that visual cues and visual acuity plays in overcoming ecological challenges to migrating birds. Other examples can be found in Rodrigo 2002 and Herbert 2017

Section 3: Rethinking Proximal and Distal Cues, Boundary Cell Information

The first proposal is informed by the idea that a rat's sense of place and the relative place of cues are mediated by its information about boundary locations in its environment. Mere sight of a cue is not enough to place it or oneself. Here is Knerim and Hamilton summarizing research on this idea.

> "In this model, the place cells do not receive a major direct input from constellations of distal cues. Rather, the influence of these cues derives from their influences on grid cells, boundary cells, and head direction cells. Thus the spatial firing of place cells does not primarily encode or represent constellations of distal cues directly. Instead, these cells represent an internally generated, spatial map that can be specific to individual contexts and can incorporate individual items and events into that framework in the support of episodic memory" (Kneirim and Hamilton 2011)

The proposal is to set aside the actual boundaries for a moment and think about "distal" and "proximal" in terms of an organism's information or representations of boundary locations. The details are informed by research on the information carried by boundary cells. Boundary cells exhibit selective firing patterns when an organism is near a boundary, like a maze wall. Meaning that their rate of electrical activity only ever jumps or dips from the base rate when an

organism is near a boundary. Each cell is tethered to a specific boundary in an organism's environment, so it only fires when the organism is near the specific boundary it is tethered to.[93] To illustrate, the boundary cell 5a recorded by Lever et al. (2009) only fires when the organism is near the North boundary of its square free-roam maze (depicted below in Fig.22). Usually, multiple boundary cells are tethered to the same boundary so that researchers observe multiple cells firing when an organism stands near a given boundary. To illustrate, cells 5a and 5e in Fig.3 each fire when the organism is near the North boundary. Boundary cells fire regardless of facing or direction, indicating that the cell activity does not depend on visual acquaintance with the boundary. They activate regardless of whether the organism approaches the boundary headfirst, sideways, or backward (Redish, 1999 & Lever et al., 2009).

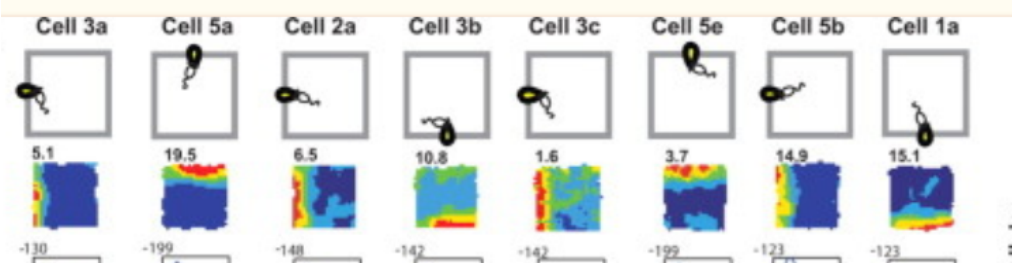


Fig.23. For each cell, the top box contains corresponding organism location and facing. Bottom box contains firing field from trials in square shaped free roam environment. Red/orange indicates interesting boundary cell activation.

[93] That is, until the cells re-map and are tethered to a new boundary when the organism learns a new environment, see Moser et al, 2015 and Jeffery, 2018.

This function means that boundary cells carry information about locations in the organism's environment that are near the actual boundaries of its environment. They only display interesting firing in those locations and return to base-rate firing everywhere else. In short, they are like boundary detectors that signal when a specific boundary is near, even when the organism cannot see the boundary. It's like having an LED that only lights up when you get near the North wall of your office. It lights up every time you're near the wall and only when you're near it, regardless of whether you see it or not. These cells indicate or detect similar information: "Heads up! *That* boundary is near!". The proposal builds on this research and tells us that the information can be used to determine a boundary separating distal and proximal cues. Cues outside 'boundary near' locations (like the red area (dark grey in greyscale version of the paper) where boundary cell 3a fires at in Fig.3) would be distal, while the cues within would be proximal.

There is significant curb appeal to these definitions. Boundary cells are thought to have the same function in natural environments, indicating when a rat is near a natural boundary like thick brush or a den wall. So, it seems like they would work in natural environments. It also fits with the working definitions from Section 1, since the maze boundaries would, ultimately, determine the locations boundary cells fire at. A researcher who thinks about the difference this way would arrive at roughly similar determinations as someone using the dominant working definitions in laboratory environments.

However, there are significant challenges to thinking about distal and proximal cues this way. First, each boundary cell's 'boundary near' information is tethered to specific locations near actual boundaries in an organism's environment. It's tethered to those locations in two ways: it's tethered in that the cells only carry information about those locations *and* in that the cells only carry that info when an organism is in those locations. A boundary cell stops carrying its 'boundary near' information when the organism moves away from the boundary the cell is tethered to, and all the cells stop carrying that information when it moves away from all the boundaries (Fig.3). This is crucial to their function as boundary detectors. Like the LED in your office, its being off when you are away from the North wall is crucial to its function as a boundary detector. It's because it only turns on when you are near the wall that it comes to carry the information "Heads up! *That* boundary is near!". If it was on all the time, it wouldn't be useful. That means that when a rat is far away from any boundaries in its environment, like when it's in the middle of a clearing in the forest (Fig.2), its boundary cells provide no 'boundary near' information about locations. It follows that there would be, in principle, no distal or proximal cues when an organism moves away from its boundaries because there would be no boundary for cues to be outside of or within. And since there would be no distal or proximal cues, those cues would not be causes or difference makers to behavior. So, the explanatory shortcomings from Section 2 that we sought to avoid pop up again when a rat moves too far from its boundaries.

Adding memory and representational abilities does not help. Even if rats had abilities to store locations the boundary cells carried the 'boundary near' information at, like if it had a cognitive map that stored allocentric spatial information about the relationship between those locations and information about whether perceived cues were beyond/within those locations (O'Keefe and Nadel, 1978 & Redish, 1999), the boundary cells would not carry that information about locations near boundaries the organism never visited, despite still have visual access to the cues beyond them. Suppose a woodrat learned to navigate a clearing by recognizing distant visual cues from its location at the center or opposite corner of the clearing. It learns to do this without visiting the boundary the cue is 'beyond', and so its boundary cells would never activate near those boundaries, giving its memory nothing to store. This would leave us unable to explain its navigation behavior in terms of distal and proximal cues.

Other problems have to do with sparse environments and long-distance travel. Environments like deserts, tundra, and even forest floors don't always have visible features that constitute boundaries. The locations boundary cells would carry information about wouldn't constitute a boundary for cues to exist beyond or within. It follows there would be no cues within or beyond those boundaries and so the distinction would fail to explain rats' navigation behaviors in those environments.

But even in environments that afford boundaries, long-distance travel behaviors like migration or hunting take rats away from the boundaries their

boundary cells are tethered to. When they travel away from home to new places, their boundary cells remain 'off' until it learns the new environment and the cells re-map (Jeffery 2018). That leaves us, again, without a boundary for separating proximal and distal cues, meaning that the distinction wouldn't contribute to explanations of behaviors where rats are away from home and, seemingly, need to rely on navigation strategies more than ever.

Section 4: Rethinking Proximal and Distal Cues, Retinal Size Information.

As we saw in the last two sections, environments seem too variable to pin definitions to boundary features, whether by pinning them to the boundary itself or a rat's information about the boundary. Once a rat moves away from the relevant boundary or enters barren environments, the distinction loses its explanatory foothold. The definitions I propose point to another source for a rat's sense of how far a visual cue is: information about the cue's effects on the rat's retinas. Visual cues affect the retinas differently based on their distance. As a rat approaches or runs away from a cue it sees, the cue leaves smaller, larger, or otherwise different 'imprints' on photoreceptors in the retinas. Information or neural representations about those changes then flow downstream to other vision centers. The proposal is to home in on some of these retinal effects of visual cues and define "distal" and "proximal" in terms of differences in the visual system's information about them. In particular, I propose definitions in terms of differences in information about a cue's retinal image size as the rat moves around its

environment. Considerations of definitions in terms of other differences in information about retinal effects are considered toward the end of the section.

A cue's retinal image size is distinct from its actual size or even its perceived size. Its retinal image size is a function of the number of photoreceptors (rods and cones) it engages or activities as a consequence of its size and distance from the photoreceptors. If we think of each photoreceptor as like a pixel on a screen, then a cue's retinal image size would be the number of pixels it activates. Increase a cue's size or move it closer to the photoreceptors and you increase its retinal image size, shrink it, or move it farther and its retinal image size shrinks too. That's different from its actual size, which is given in measures like feet or inches and remains constant whether it's closer or farther from photoreceptors. It's also different from a cue's perceived size which is how its size appears to someone who is conscious of it.

To illustrate the different kinds of sizes, consider the orange (light grey) discs from the Ebbinghaus illusion in Fig.24. The retinal image size of the discs is the same (or roughly similar) because they activate the same number of photoreceptors. That's because, despite appearances, the discs are the same actual size and distance away. They create the same visual angle and cast similar projections on the photoreceptors or pixels of the retina. The perceived size, however, is different between the discs, which is the point of the illusion. Our automatic visual processes mistake the outer rings of discs for depth cues and deliver conscious perception of different size discs on a 2D plane. So, the

perceived size of the discs is different while the retinal image size is the same. Its actual size can be given in inches or screen pixels.

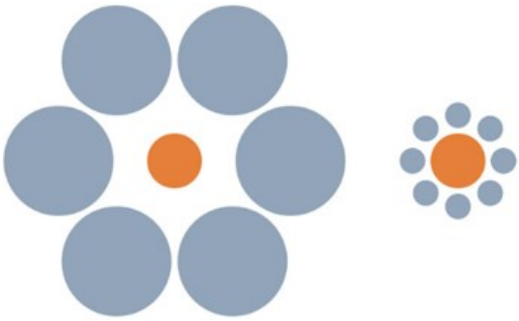


Fig.24 Ebbinghaus Illusion. Thomson and Macpherson, 2017

In mammals like rodents, apes, and humans, a cue's retinal image size is collected by photoreceptors in the retina and information is transferred to later, upstream visual processing areas via relays leading from the retina to the lateral geniculate nucleus to visual processing centers in the visual cortex (Palmer, 1999). I'm suggesting that researchers think about the difference between distal and proximal cues in terms of that information contribution. If there is no difference in the information a rat has about a cue's retinal image size as the organism moves around its environment, then the cue is distal. In other words, if information about the cue's 'pixel size' on the 'retina screen' stays the same as the organism scurries, hunts, or explores, then the cue should be counted as distal. If, on the other hand, there are changes to information about retinal image size, or the cue's pixel size, as it moves about, then the cue is proximal. So, the relevant question stops being about where the cue is relative to some boundary and

becomes a question about a syntax feature of the information about retinal image size in early vision: does that information change or stay the same as the organism moves about? To put the proposal in terms of definitions, distal or far-off visual cues are defined as cues that result in stable, unchanging retinal image size information as an organism moves around and proximal or near visual cues result in changing information.

Consider two mechanisms by which information about a cue's retinal image size could remain constant while a rat moves about. The first has to do with detection of the cue by photoreceptors in the retina. While some photoreceptors, like the cones responsible for color detection can be sensitive to even one photon changes of light, the rod receptors responsible for retinal size detection require a threshold of change to activate. The detectors are specialized so that the chemical process by which they activate or 'kick off' are not sensitive to every nuance of difference in stimuli.[94] They are only sensitive to those that meet a threshold of change. Information about the arrays of activate and inactivate photoreceptors across the retina is carried down the vision pathway by relays that deliver that information, without processing, to early vision processing centers in the visual cortex or motor centers responsible for pupil and head movement (Van den Bergh et al 2010 and Laramee and Boire, 2015).[95]

---

[94] For discussions of specialized differences between nocturnal and diurnal animals see Ross & Kirk, 2006. For reviews/studies of how the cells become more specialized during covert attention, locomotion or other affective states in mice, rats, and humans see Ferguson and Cardin, 2020, Foster et al 2020, Neske et al 2019, Jurjut et al 2017, and Soma et al 2012.

[95] For overviews of the visual pathway in healthy and injured or sick rats see Dean 1981, Thuen, et al 2005, Usrey and Alitto 2015. Some research suggests a little 'pruning' or cleaning of the retinal

This mechanism would lead to stable retinal image size information if the light changes that covary with changes in distance and visual angle as an organism moves about its environment are too small to be detected by photoreceptors. Like if the changes to a distant mountain peak's visual angle are too small to be detected by the photoreceptors of a rat scurrying around its corner of the forest. The idea is that the rat's eye would never record the change in retinal image size and its input to early or later vision would be the same, despite the rat's movement. The retinal image size would be recorded by the retina and information about the recording would be relayed to other vision or motor centers, becoming an important input for processing or behavior.

The second has to do with top-down processing on the retinal size information stored in early vision processing centers like the V1. Studies have demonstrated that retinal size information can be modulated by attention, locomotion, or even information about perceived size (Flossman et al 2021, Froudarakis et al 2020 and Zeng et al 2020). Processing mechanisms, like the ones responsible for the Ebbinghaus illusion, could modulate the retinal image size info so it stays the same as a consequence of the cue's relation to other cues, familiarity, or salience. So, even though the eye would pick up on the changes

image size takes place in the lateral geniculate nucleus (Tang et al. 2016 and Weyand, 2015). This suggests more thresholds where changes to visual angle may not be strong enough to meet a threshold for being detected, and so information about retinal image size would remain the same while an organism moved around.

because they meet the threshold, upstream visual processes would tamper down and hold the information steady before it is taken up in visual processing.

Either mechanism leads to stable, unchanging retinal image size information while an organism moves about its environment. On the first mechanism, the stimulus differences that lead to differences in the visual angle that are detected by photoreceptors are too small and do not meet the threshold for detection. On the second, top-down processing mechanisms hold the information fixed in early vision.

Thinking about the distinction along these lines fits with scientific thinking in two, important ways. First, it preserves scientific thinking about which cues are distal and which are proximal in Morris water maze studies and other laboratory environments. Or, to put the point philosophically, it preserves the extension of these scientific categories in good cases. In experiments involving distal and proximal cues, researchers place the cues they think of as distal far outside the water maze boundaries in order to avoid confounding the effects of those cues with the effects of proximal cues within the boundaries. Because of the distance, changes to a cue's visual angle are more likely to go undetected or to be compensated for by top-down processing, yielding a stable retinal image size as the organism moves around. On the other hand, cues thought of as proximal are usually located closer to the organism, within the water maze boundaries. Changes to their visual retinal image size are much more likely to be detected or

affected by top-down processing as an organism moves around.[96] So, there's not much change in which cues count as distal and which cues count as proximal, what changes is *why* we think of those cues as distal or proximal.

Second, it helps us understand the explanatory roles assigned to proximal and distal cues. As explained in Section 1, researchers draw the distinction because, in study after study, rats use the cues differently. They use the cues beyond the maze preferentially and for direction or orientation information (relevant to questions like "which direction should I head in?"). This way of thinking helps us understand the connection between the cues and the differences in information; it explains those explanatory roles.[97]

Cues that produce stable retinal size information are usually large, far away, and insensitive to the rat's local going-ons, like wind or rain. The size and distance make them reliable, which explains why rats use them preferentially. A rat that needs to make a quick, high-stakes decision about where to scurry will make a bet on the more reliable cue. It also explains why rats do not use distal cues for distance information and rely on proximal cues instead. A rat's information about its distance from a cue or the distance between two cues is determined by calculating or associating information about changes to retinal

[96] While our abilities to neatly categorize get fuzzy in experiments where cues are placed closer to apparatus boundaries or where apparatus boundaries can be far away, that is consistent with the way researchers think about things on the standard way. In their characterization, Kneirem and Hamilton (2011) explain that cases like these present a grey area for the dominant working definitions.

[97] Contrast that with the dominant working definitions, which cannot even assign explanatory roles to those cues in natural environments. We can think of this as like an added bonus.

image size with information about motor function (Kneierim & Hamilton, 2011 & Jeffery, 2018). It's a function of the rate at which its retinal image size changes in proportion to the rate at which a rat approaches or retreats from the cue. The stability of retinal image size information makes distal cues bad sources of direction information because there is no change in the retinal image size to calculate and associate with changes to information about motor functions. From the perspective of retinal image size, they're always the same distance away.

This way of thinking also avoids the problem I raised in Section 2 because it promises to make the distinction explanatory of behaviors in natural environments. What makes visual cues like a tree, coastline, or mountain peak distal/far away rather than proximal/near? Why predict that a rat will use it to change direction rather than gauge its distance? We can give an answer in natural environments and say it is because the cue(s) affects a rat's retinas in such a way that the information about those cues is different. Unlike the working definitions from Section 1, there's nothing about natural environments that prevents the terms, defined this way, from being explanatory in those environments.

I mentioned at the beginning of Section 4 that definitions could also be developed around differences in information about other retinal effects besides changes to retinal image size. Research on spatial cognition identifies parallax effects from motion or stereoscopic vision as important sources of information

about a cue's size or distance.[98] It seems researchers could use these or other relevant effects for definitions instead of developing them around retinal image size. This prompts questions like "why define "distal"/"proximal" in terms of information about retinal image size and not in terms of information about other effects like motion parallax effects?"

I think an answer depends on what researchers are trying to explain. The cues that produce stable information about retinal image size will also usually produce stable or relatively reduced parallax effects from motion because the cues are far enough away from the viewing point(s). What counts as distal/proximal won't change much, if at all, between definitions. But we might take information about, say, motion parallax effects to be more important to an organism based on what else we know about the organism. It may appear obvious to biologists that an organism is using motion parallax effect information to gauge a cue's distance, like if they saw a rat (or other organism like a meercat) bobbing its head up and down a lot before moving. It makes sense for researchers to define "distal"/"proximal" in terms of information about the motion parallax effects in those cases. Otherwise, I do not see a reason for preferring one set of definitions over the other if they deliver the same extension, if they deliver the same verdict about which cues are proximal and which are distal.

[98] See, for instance Hok, Oucet, Duvelle, Save, and Sargolini 2016. For research on how even small, microscopic head movements-the kind most organisms make all the time to keep balance-can have detectable parallax effects see Aytekin and Rucci 2012.

To summarize, the proposal I offered identifies differences in information about retinal effects as the relevant difference for defining distal and proximal visual cues in Morris water maze studies. In particular, it identifies differences between information about a cue's retinal image size as the rat moves around. Distal visual cues are defined as cues that result in stable, unchanging retinal image size information as an organism moves around and proximal cues are ones that result in changing information. The definitions promise to work in natural environments and fit with research done in laboratories. There may be other similar definitions that usually converge in extension. I argued that the definitions researchers choose is a matter of context and depends on what researchers are interested in explaining or predicting.

Conclusion

Rats use visual cues differently as they navigate their environments. One distinction researchers use to characterize differences between cues is the distinction between distal and proximal cues. The dominant working definitions of "distal" and "proximal" involve thinking about distal cues as beyond an experimental apparatus and proximal cues as within that apparatus. I argued that there is a problem with thinking about cues this way, and recommended a new way of thinking about the distinction in terms of the information rats have about a cue's retinal image size.

Suppose my arguments are successful. What does that mean for the relevant scientific paradigms and their claims about the explanatory roles proximal and distal visual cues? It means there are, at least, two sets of definitions for "distal"/ "proximal", two things a researcher could mean by "distal" when they say things like "rats rely heavily on distal visual cues for navigation". They could mean *that thing outside the maze boundary* or they could mean *that thing that produces a stable or relatively stable retinal image size representation in early vision as the organism moves around*. Context will have a lot to do with it. If they are predicting woodrat behavior in the woods, they must mean the latter. If they are explaining why rats deviate swim paths in a Morris water maze, they likely, but not necessarily, mean the former (it is just so practical).

Bibliography

Alyan, S., Touretzky, D., & Taube, J. (1995). The Involvement of Passive Path Integration in Learning the Morris Water Maze. *Social Neuroscience*, *21*.

Andrews, Kristin and Susana Monsó, "Animal Cognition", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2021/entries/cognition-animal

Arias-Cavieres, A., Rozas, C., Reyes-Parada, M., Barrera, N., Pancetti, F., Loyola, S., Lorca, R. A., Zeise, M. L., & Morales, B. (2010). MDMA ("ecstasy") impairs learning in the Morris Water Maze and reduces hippocampal LTP in young rats. *Neuroscience Letters*, *469*(3), 375–379. https://doi.org/10.1016/j.neulet.2009.12.031

Ayer, A. J. (1970). *Language, truth and logic* (Unabridged and unaltered republ. of the 2. (1946) ed). Dover Publications.

Baars, B. J., Geld, N., & Kozma, R. (2021). Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments. *Frontiers in Psychology*, *12*, 749868. https://doi.org/10.3389/fpsyg.2021.749868

Beltran, R. S., Yuen, A. L., Condit, R., Robinson, P. W., Czapanskiy, M. F., Crocker, D. E., & Costa, D. P. (2022). Elephant seals time their long-distance migrations using a map sense. *Current Biology*, *32*(4), R156–R157. https://doi.org/10.1016/j.cub.2022.01.031

Benhamou, S. (1997). Path integration by swimming rats. *Animal Behaviour*, *54*(2), 321–327. https://doi.org/10.1006/anbe.1996.0464

Biegler, R. (2000). Possible uses of path integration in animal navigation. *Animal Learning & Behavior*, *28*(3), 257–277. https://doi.org/10.3758/BF03200260

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Block, N. J. (2022). *The border between seeing and thinking*. Oxford University Press.

Bohannon, R. W., & Endemann, N. (1989). How accurately can elbow flexion force be estimated? *Perceptual and Motor Skills*, *68*(3 Pt 2), 1159–1162. https://doi.org/10.2466/pms.1989.68.3c.1159

Bostock, E., Muller, R. U., & Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, *1*(2), 193–205. https://doi.org/10.1002/hipo.450010207

Brandeis, R., Brandys, Y., & Yehuda, S. (1989). The Use of the Morris Water Maze in the Study of Memory and Learning. *International Journal of Neuroscience*, *48*(1–2), 29–69. https://doi.org/10.3109/00207458909002151

Broadbent, N. J., Squire, L. R., & Clark, R. E. (2006). Reversible hippocampal lesions disrupt water maze performance during both recent and remote memory tests. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *13*(2), 187–191. https://doi.org/10.1101/lm.134706

Burge, T. (2010). *Origins of objectivity*. Oxford University Press.

Camp, E. (2018). *Why Maps are Not Propositional* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198732570.003.0002

Campbell, J., & Cassam, Q. (2014). *Berkeley's puzzle: What does experience teach us?* (First edition). Oxford University Press.

Carman, H. (2002). Proximal versus Distal Cue Utilization in Spatial Navigation: The Role of Visual Acuity?,. *Neurobiology of Learning and Memory*, *78*(2), 332–346. https://doi.org/10.1006/nlme.2002.4062

Carr, H., & Watson, J. B. (1908). Orientation in the white rat. *Journal of Comparative Neurology and Psychology*, *18*(1), 27–44. https://doi.org/10.1002/cne.920180103

Chapillon, P., & Roullet, P. (1996). Use of proximal and distal cues in place navigation by mice changes during ontogeny. *Developmental Psychobiology*, *29*(6), 529–545. https://doi.org/10.1002/(SICI)1098-2302(199609)29:6<529::AID-DEV5>3.0.CO;2-O

Chiba, Andrea. (2015) "Why we use rodents to research the brain" in The San Diego Union Tribune. https://www.sandiegouniontribune.com/news/science/sdut-brain-logic-behind-using-rodents-2015apr02-htmlstory.html

Condit, R., Beltran, R. S., Robinson, P. W., Crocker, D. E., & Costa, D. P. (2022). Birth timing after the long feeding migration in northern elephant seals. *Marine Mammal Science*, *38*(3), 931–940. https://doi.org/10.1111/mms.12896

Craig, S., Cunningham, L., Kelly, L., & Commins, S. (2005). Long-term retention and overshadowing of proximal and distal cues following habituation in an object

exploration task. *Behavioural Processes*, *68*(2), 117–128. https://doi.org/10.1016/j.beproc.2004.12.001

Crystal, J. D. (2012). Validating animal models of metacognition. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), Foundations of metacognition (pp. 36–49). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199646739.003.0003

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, *72*(20), 741. https://doi.org/10.2307/2024640

Darwin, C.R. (1873). The Origin of Certain Instincts. *Nature*, *7*, 417–418.

Davidson, D. (1968). On saying that. *Synthese*, *19*(1–2), 130–146. https://doi.org/10.1007/BF00568054

Davies, P. S. (2000). Malfunctions. *Biology and Philosophy*, *15*(1), 19–38.

Day, L. B., & Schallert, T. (1996). Anticholinergic effects on acquisition of place learning in the Morris water task: Spatial mapping deficit or inability to inhibit nonplace strategies? *Behavioral Neuroscience*, *110*(5), 998–1005. https://doi.org/10.1037//0735-7044.110.5.998

Dean, P. (1981). Visual pathways and acuity in hooded rats. *Behavioural Brain Research*, *3*(2), 239–271. https://doi.org/10.1016/0166-4328(81)90050-4

Dean, P., Redgrave, P., & Westby, G. W. (1989). Event or emergency? Two response systems in the mammalian superior colliculus. *Trends in Neurosciences*, *12*(4), 137–147. https://doi.org/10.1016/0166-2236(89)90052-0

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529–14534. https://doi.org/10.1073/pnas.95.24.14529

Dennett, D. C. (2010). *Content and consciousness*. Routledge.

Dickie, I. (2020). Cognitive Focus. In I. Dickie, *Singular Thought and Mental Files* (pp. 223–250). Oxford University Press. https://doi.org/10.1093/oso/9780198746881.003.0011

Drayson, Z. (2014). The Personal/Subpersonal Distinction. *Philosophy Compass*, *9*(5), 338–346.

Dretske, F. I. (1999). *Knowledge and the flow of information*. CSLI Publications.

Ferguson, K. A., & Cardin, J. A. (2020). Mechanisms underlying gain modulation in the cortex. *Nature Reviews. Neuroscience*, *21*(2), 80–92. https://doi.org/10.1038/s41583-019-0253-y

Fitting, S., Allen, G. L., & Wedell, D. H. (2007). Remembering Places in Space: A Human Analog Study of the Morris Water Maze. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), *Spatial Cognition V Reasoning, Action, Interaction* (Vol. 4387, pp. 59–75). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75666-8_4

Flossmann, T., & Rochefort, N. L. (2021). Spatial navigation signals in rodent visual cortex. *Current Opinion in Neurobiology*, *67*, 163–173. https://doi.org/10.1016/j.conb.2020.11.004

Fodor, J. A. (1985). Précis of *The Modularity of Mind*. *Behavioral and Brain Sciences*, *8*(1), 1–5. https://doi.org/10.1017/S0140525X0001921X

Fodor, J. A. (2008). *The modularity of mind: An essay on faculty psychology* (15. printing). MIT Press.

Foster, J. J., Thyer, W., Wennberg, J. W., & Awh, E. (2021). Covert Attention Increases the Gain of Stimulus-Evoked Population Codes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *41*(8), 1802–1815. https://doi.org/10.1523/JNEUROSCI.2186-20.2020

Francis, D. (1995). Stress-induced disturbances in Morris water-maze performance: Interstrain variability. *Physiology & Behavior*, *58*(1), 57–65. https://doi.org/10.1016/0031-9384(95)00009-8

Frege, G. (1948). Sense and Reference. *The Philosophical Review*, *57*(3), 209. https://doi.org/10.2307/2181485

Frost, B. J., & Mouritsen, H. (2006). The neural mechanisms of long distance animal navigation. *Current Opinion in Neurobiology*, *16*(4), 481–488. https://doi.org/10.1016/j.conb.2006.06.005

Gallistel, C. R. (1993). *The Organization of learning* (1. MIT press paperb. ed). MIT Press.

Garson, J. (2007). Function and Teleology. In S. Sahotra & A. Plutynski (Eds.), *A Companion to the Philosophy of Biology* (pp. 525–549). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470696590.ch28

Garson, J. (2011). Selected effects and causal role functions in the brain: The case for an etiological approach to neuroscience. *Biology & Philosophy*, *26*(4), 547–565. https://doi.org/10.1007/s10539-011-9262-6

Garson, Justin. *A critical overview of biological functions*. (2016). Springer Berlin Heidelberg.

Garthe, A., & Kempermann, G. (2013). An old test for new neurons: Refining the Morris water maze to study the functional relevance of adult hippocampal neurogenesis. *Frontiers in Neuroscience*, *7*, 63. https://doi.org/10.3389/fnins.2013.00063

Gaudio, J. L., & Snowdon, C. T. (2008). Spatial cues more salient than color cues in cotton-top tamarins (Saguinus oedipus) reversal learning. *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *122*(4), 441–444. https://doi.org/10.1037/0735-7036.122.4.441

Gehring, T. V., Luksys, G., Sandi, C., & Vasilaki, E. (2015a). Detailed classification of swimming paths in the Morris Water Maze: Multiple strategies within one trial. *Scientific Reports*, *5*(1), 14562. https://doi.org/10.1038/srep14562

Gehring, T. V., Luksys, G., Sandi, C., & Vasilaki, E. (2015b). Detailed classification of swimming paths in the Morris Water Maze: Multiple strategies within one trial. *Scientific Reports*, *5*(1), 14562. https://doi.org/10.1038/srep14562

Georgopoulos, A. P., Caminiti, R., Kalaska, J. F., & Massey, J. T. (1983). Spatial Coding of Movement: A Hypothesis Concerning the Coding of Movement Direction by Motor Cortical Populations. In J. Massion, J. Paillard, W. Schultz, & M. Wiesendanger (Eds.), *Neural Coding of Motor Performance* (Vol. 7, pp. 327–336). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-68915-4_34

Glüer, K. (2014). Looks, Reasons, and Experiences. In B. Brogaard (Ed.), *Does Perception Have Content?* (pp. 76–102). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199756018.003.0004

Godfrey-Smith, P. (1993). FUNCTIONS: CONSENSUS WITHOUT UNITY. *Pacific Philosophical Quarterly*, *74*(3), 196–208. https://doi.org/10.1111/j.1468-0114.1993.tb00358.x

Godfrey-Smith, P. (2012). Signals, Icons, and Beliefs. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and Her Critics* (pp. 41–62). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118328118.ch2

Grieves, R. M., Jedidi-Ayoub, S., Mishchanchuk, K., Liu, A., Renaudineau, S., & Jeffery, K. J. (2020). The place-cell representation of volumetric space in rats. *Nature Communications*, *11*(1), 789. https://doi.org/10.1038/s41467-020-14611-7

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harrison, F. E., Reiserer, R. S., Tomarken, A. J., & McDonald, M. P. (2006). Spatial and nonspatial escape strategies in the Barnes maze. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *13*(6), 809–819. https://doi.org/10.1101/lm.334306

Hebb, D. O. (1938). The innate organization of visual activity. III. Discrimination of brightness after removal of the striate cortex in the rat. *Journal of Comparative Psychology*, *25*(2), 427–437. https://doi.org/10.1037/h0060221

Hébert, M., Bulla, J., Vivien, D., & Agin, V. (2017). Are Distal and Proximal Visual Cues Equally Important during Spatial Learning in Mice? A Pilot Study of Overshadowing in the Spatial Domain. *Frontiers in Behavioral Neuroscience*, *11*, 109. https://doi.org/10.3389/fnbeh.2017.00109

Heck, R.-K. (2007). Are there different kinds of content? In B. P. McLaughlin & J. D. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 117–138). Blackwell.

Honzick, Charles. (1936). *The Sensory Basis of Maze Learning in Rats*.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huberman, A. D., & Niell, C. M. (2011). What can mice tell us about how vision works? *Trends in Neurosciences*, *34*(9), 464–473. https://doi.org/10.1016/j.tins.2011.07.002

Ingram, N. T., Sampath, A. P., & Fain, G. L. (2016). Why are rods more sensitive than cones? *The Journal of Physiology*, *594*(19), 5415–5426. https://doi.org/10.1113/JP272556

Jeffery, K. J. (2011). Place Cells, Grid Cells, Attractors, and Remapping. *Neural Plasticity*, *2011*, 1–11. https://doi.org/10.1155/2011/182602

Jeffery, K. J. (2018). Cognitive representations of spatial location. *Brain and Neuroscience Advances*, *2*, 2398212818810686. https://doi.org/10.1177/2398212818810686

Jeshion, R. (2010). Singular thought: Acquaintance, semantic instrumentalism, and cognitivism. In R. Jeshion (Ed.), *New Essays on Singular Thought* (pp. 105--141). Oxford University Press.

Jones, Keri. (2022). *Morris Water Maze*. https://www.augusta.edu/research/core/sabc/test-spatial-learning-memory.php

Jurjut, O., Georgieva, P., Busse, L., & Katzner, S. (2017). Learning Enhances Sensory Processing in Mouse V1 before Improving Behavior. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(27), 6460–6474. https://doi.org/10.1523/JNEUROSCI.3485-16.2017

Kant, I., & Reath, A. (1997). *Immanuel Kant Critique of Practical Reason:* (M. J. Gregor, Ed.; 1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511809576

Kasinos, N., Lilley, G. A., Subbarao, N., & Haneef, I. (1992). A robust and efficient automated docking algorithm for molecular recognition. *Protein Engineering*, *5*(1), 69–75. https://doi.org/10.1093/protein/5.1.69

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.

King, J. C. (2020). Singular Thought, Russellianism, and Mental Files. In J. C. King, *Singular Thought and Mental Files* (pp. 73–106). Oxford University Press. https://doi.org/10.1093/oso/9780198746881.003.0005

Kitcher, P. (2015). Experimental Animals: Experimental Animals. *Philosophy & Public Affairs*, *43*(4), 287–311. https://doi.org/10.1111/papa.12060

Knierim, J. J., & Hamilton, D. A. (2011). Framing Spatial Cognition: Neural Representations of Proximal and Distal Frames of Reference and Their Roles in Navigation. *Physiological Reviews*, *91*(4), 1245–1279. https://doi.org/10.1152/physrev.00021.2010

Laramée, M.-E., & Boire, D. (2014). Visual cortical areas of the mouse: Comparison of parcellation and network structure with primates. *Frontiers in Neural Circuits*, *8*, 149. https://doi.org/10.3389/fncir.2014.00149

Lesley A. Schimanski, Peter Lipa, & Carol A. Barnes. (2013). Tracking the Course of Hippocampal Representations during Learning: When Is the Map Required? *The Journal of Neuroscience*, *33*(7), 3094. https://doi.org/10.1523/JNEUROSCI.1348-12.2013

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary Vector Cells in the Subiculum of the Hippocampal Formation. *Journal of Neuroscience*, *29*(31), 9771–9777. https://doi.org/10.1523/JNEUROSCI.1319-09.2009

Lewis, D. K. (1969). *Convention: A Philosophical Study* (Vol. 20, Issue 80, p. 286). Wiley-Blackwell.

Lewis, D. K. (2001). *On the plurality of worlds*. Blackwell Publishers.

Ludvig, N. (1999). Place cells can flexibly terminate and develop their spatial firing. A new theory for their function. *Physiology & Behavior*, *67*(1), 57–67. https://doi.org/10.1016/s0031-9384(99)00048-7

Macpherson, F. (2013). The Philosophy and Psychology of Hallucination: An Introduction. In F. Macpherson & D. Platchias (Eds.), *Hallucination* (pp. 1–38). The MIT Press. https://doi.org/10.7551/mitpress/9780262019200.003.0001

Matsumura, M., Watanabe, Y. Y., Robinson, P. W., Miller, P. J. O., Costa, D. P., & Miyazaki, N. (2011). Underwater and surface behavior of homing juvenile northern elephant seals. *Journal of Experimental Biology*, *214*(4), 629–636. https://doi.org/10.1242/jeb.048827

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. https://doi.org/10.1037/0033-295X.102.3.419

McDowell, J. H. (1996). *Mind and world: With a new introduction* (1st Harvard University Press paperback ed). Harvard University Press.

McGauran, A.-M. T., Harvey, D., Cunningham, L., Craig, S., & Commins, S. (2004). Retention of cue-based associations in the water maze is time-dependent and sensitive to disruption by rotating the starting position. *Behavioural Brain Research*, *151*(1–2), 255–266. https://doi.org/10.1016/j.bbr.2003.09.005

Means, L. W., Alexander, S. R., & O'Neal, M. F. (1992). Those cheating rats: Male and female rats use odor trails in a water-escape "working memory" task. *Behavioral and Neural Biology*, *58*(2), 144–151. https://doi.org/10.1016/0163-1047(92)90387-J

Melcón, M., Stern, E., Kessel, D., Arana, L., Poch, C., Campo, P., & Capilla, A. (2023). *Perception of near-threshold visual stimuli is influenced by pre-stimulus*

alpha-band amplitude but not by alpha phase [Preprint]. Neuroscience. https://doi.org/10.1101/2023.03.14.532551

Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, *86*(6), 281--297.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Vol. 14, Issue 1, pp. 51–56). MIT Press.

Millikan, R. G. (1999). [No title found]. *Philosophical Studies*, *95*(1/2), 45–65. https://doi.org/10.1023/A:1004532016219

Millikan, R. G. (2012). ARE THERE MENTAL INDEXICALS AND DEMONSTRATIVES?: *Are There Mental Indexicals and Demonstratives? Philosophical Perspectives*, *26*(1), 217–234. https://doi.org/10.1111/phpe.12004

Mittelstaedt, M.-L., & Mittelstaedt, H. (1980). Homing by path integration in a mammal. *Naturwissenschaften*, *67*(11), 566–567. https://doi.org/10.1007/BF00450672

Mogensen, J. (2011). Reorganization of the Injured Brain: Implications for Studies of the Neural Substrate of Cognition. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00007

Mogensen, J., & Malá, H. (2009). Post-traumatic functional recovery and reorganization in animal models: A theoretical and methodological challenge. *Scandinavian Journal of Psychology*, *50*(6), 561–573. https://doi.org/10.1111/j.1467-9450.2009.00781.x

Morgan, A. (2018). Mindless accuracy: On the ubiquity of content in nature. *Synthese*, *195*(12), 5403–5429. https://doi.org/10.1007/s11229-018-02011-w

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*(1), 47–60. https://doi.org/10.1016/0165-0270(84)90007-4

Nagahara, A. H., Otto, T., & Gallagher, M. (1995). Entorhinal-perirhinal lesions impair performance of rats on two versions of place learning in the Morris water maze. *Behavioral Neuroscience*, *109*(1), 3–9. https://doi.org/10.1037/0735-7044.109.1.3

Neander, K. (1991). Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science*, *58*(2), 168–184.

Neander, K. (1995). Misrepresenting and malfunctioning. *Philosophical Studies*, *79*(2), 109–141.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT Press.

Neske, G. T., Nestvogel, D., Steffan, P. J., & McCormick, D. A. (2019). Distinct Waking States for Strong Evoked Responses in Primary Visual Cortex and Optimal Visual Detection Performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *39*(50), 10044–10059. https://doi.org/10.1523/JNEUROSCI.1226-18.2019

Niikawa, T. (2023). Naïve realism, imagination and hallucination. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-023-09915-0

Noë, A. (2004). *Action in perception*. MIT press.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, *51*(1), 78–109. https://doi.org/10.1016/0014-4886(76)90055-8

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

O'Keefe, J., & Krupic, J. (2021). Do hippocampal pyramidal cells respond to nonspatial stimuli? *Physiological Reviews*, *101*(3), 1427–1456. https://doi.org/10.1152/physrev.00014.2020

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press.

Oostra, B. A., & Nelson, D. L. (2006). Animal Models of Fragile X Syndrome: Mice and Flies. In *Genetic Instabilities and Neurological Diseases* (pp. 175–193). Elsevier. https://doi.org/10.1016/B978-012369462-1/50012-0

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Oxford University Press.

Othman, M. Z., Hassan, Z., & Che Has, A. T. (2022). Morris water maze: A versatile and pertinent tool for assessing spatial learning and memory. *Experimental Animals*, *71*(3), 264–280. https://doi.org/10.1538/expanim.21-0120

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.

Park, M., Kim, C.-H., Jo, S., Kim, E. J., Rhim, H., Lee, C. J., Kim, J. J., & Cho, J. (2015). Chronic Stress Alters Spatial Representation and Bursting Patterns of

Place Cells in Behaving Mice. *Scientific Reports*, *5*(1), 16235. https://doi.org/10.1038/srep16235

Parron, C., Poucet, B., & Save, E. (2004). Entorhinal cortex lesions impair the use of distal but not proximal landmarks during place navigation in the rat. *Behavioural Brain Research*, *154*(2), 345–352. https://doi.org/10.1016/j.bbr.2004.03.006

Peirce, C. S., Hartshorne, C., Weiss, P., & Peirce, C. S. (1985). *Principles of philosophy: Two volumes in one* (5. [printing]). Belknap Press of Harvard Univ. Press.

Perry, J. (2020). Singular Thoughts. In J. Perry, *Singular Thought and Mental Files* (pp. 143–158). Oxford University Press. https://doi.org/10.1093/oso/9780198746881.003.0007

Plantinga, A. (1993). *Warrant and proper function*. Oxford University Press.

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book* (First edition). Oxford University Press.

Poulter, S., Hartley, T., & Lever, C. (2018). The Neurobiology of Mammalian Navigation. *Current Biology: CB*, *28*(17), R1023–R1042. https://doi.org/10.1016/j.cub.2018.05.050

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. https://doi.org/10.1016/0010-0277(89)90014-0

Quirk, G. J., Muller, R. U., & Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *10*(6), 2008–2017. https://doi.org/10.1523/JNEUROSCI.10-06-02008.1990

Ramos, J. M. J. (1998). Retrograde amnesia for spatial information: A dissociation between intra and extramaze cues following hippocampus lesions in rats: Retrograde amnesia for spatial information. *European Journal of Neuroscience*, *10*(10), 3295–3301. https://doi.org/10.1046/j.1460-9568.1998.00388.x

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Ranck, J. B. (1973). Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats. *Experimental Neurology*, *41*(2), 462–531. https://doi.org/10.1016/0014-4886(73)90290-2

Redish, A. D. (1999). *Beyond the cognitive map: From place cells to episodic memory*. MIT Press.

Redish, A. D., & Touretzky, D. S. (1998). The Role of the Hippocampus in Solving the Morris Water Maze. *Neural Computation*, *10*(1), 73–111. https://doi.org/10.1162/089976698300017908

Reimer, M. (2020). Descriptive Names and Singular Thought: Reflections on the Evans/Kaplan Debate. In M. Reimer, *Singular Thought and Mental Files* (pp. 38–51). Oxford University Press. https://doi.org/10.1093/oso/9780198746881.003.0003

Rescorla, M. (2009). Cognitive maps and the language of thought. *British Journal for the Philosophy of Science*, *60*(2), 377–407.

Rivard, B., Li, Y., Lenck-Santini, P.-P., Poucet, B., & Muller, R. U. (2004). Representation of objects in space by two classes of hippocampal pyramidal cells. *The Journal of General Physiology*, *124*(1), 9–25. https://doi.org/10.1085/jgp.200409015

Rodrigo, T. (2002). Navigational Strategies and Models. *Psicológica*.

Rodriguez, P. F. (2010). Human navigation that requires calculating heading vectors recruits parietal cortex in a virtual and visually sparse water maze task in fMRI. *Behavioral Neuroscience*, *124*(4), 532–540. https://doi.org/10.1037/a0020231

Ross, C. F., & Kirk, E. C. (2007). Evolution of eye size and shape in primates. *Journal of Human Evolution*, *52*(3), 294–313. https://doi.org/10.1016/j.jhevol.2006.09.006

Rupert, R. D. (2011). Embodiment, Consciousness, and the Massively Representational Mind. *Philosophical Topics*, *39*(1), 99–120. JSTOR.

Russell, B. (1905). On Denoting. *Mind*, *14*(56), 479–493. JSTOR.

Russell, B. (1910). Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society*, *11*, 108–128. JSTOR.

Schallert, T. (2006). Behavioral tests for preclinical intervention assessment. *NeuroRX*, *3*(4), 497–504. https://doi.org/10.1016/j.nurx.2006.08.001

Schellenberg, S. (2006). Sellarsian Perspectives on Perception and Non-Conceptual Content. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *92*(1), 173–196.

Schellenberg, S. (2018). *The unity of perception: Content, consciousness, evidence* (First edition). Oxford University Press.

Schoenfeld, R., Schiffelholz, T., Beyer, C., Leplow, B., & Foreman, N. (2017). Variants of the Morris water maze task to comparatively assess human and rodent place navigation. *Neurobiology of Learning and Memory*, *139*, 117–127. https://doi.org/10.1016/j.nlm.2016.12.022

Seguinot, V., Cattet, J., & Benhamou, S. (1998). Path integration in dogs. *Animal Behaviour*, *55*(4), 787–797. https://doi.org/10.1006/anbe.1997.0662

Shusterman, R. (2008). *Body consciousness: A philosophy of mindfulness and somaesthetics*. Cambridge University Press.

Siegel, S. (2010). *The Contents of Visual Experience*. Oxford University Press USA.

Siegel, S. (2017). *The Rationality of Perception*. Oxford University Press.

Siegel, Susanna. (2021). The Contents of Perception. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=perception-contents

Singer, P. (1995). *Animal liberation* (2. ed., with a new preface by the author).

Pimlico.

Skyrms, B. (1995). Strict coherence, sigma coherence and the metaphysics of quantity. *Philosophical Studies*, *77*(1), 39–55.

Snowdon, P. (2008). Hinton and the Origins of Disjunctivism. In A. Haddock & F. Macpherson (Eds.), *Disjunctivism* (1st ed., pp. 35–56). Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780199231546.003.0002

Soma, S., Shimegi, S., Suematsu, N., & Sato, H. (2013). Cholinergic modulation of response gain in the rat primary visual cortex. *Scientific Reports*, *3*, 1138. https://doi.org/10.1038/srep01138

Stackman, R. W., Lora, J. C., & Williams, S. B. (2012). Directional responding of C57BL/6J mice in the Morris water maze is influenced by visual and vestibular cues and is dependent on the anterior thalamic nuclei. *The Journal of*

Neuroscience: The Official Journal of the Society for Neuroscience, *32*(30), 10211–10225. https://doi.org/10.1523/JNEUROSCI.4868-11.2012

Stampe, D. W. & Philosophy Documentation Center. (1977). Toward a Causal Theory of Linguistic Representation. *Midwest Studies in Philosophy*, *2*, 42–63. https://doi.org/10.1111/j.1475-4975.1977.tb00027.x

Stebbing, L. S. (2018). *A modern introduction to logic*. Routledge.

Sullivan, J. A. (2010). Reconsidering 'spatial memory' and the Morris water maze. *Synthese*, *177*(2), 261–283. https://doi.org/10.1007/s11229-010-9849-5

Sullivan-Bissett, E. (2017). Malfunction defended. *Synthese*, *194*(7), 2501–2522. https://doi.org/10.1007/s11229-016-1062-8

Suzuki, S., Augerinos, G., & Black, A. H. (1980). Stimulus control of spatial behavior on the eight-arm maze in rats. *Learning and Motivation*, *11*(1), 1–18. https://doi.org/10.1016/0023-9690(80)90018-1

Tang, J., Ardila Jimenez, S. C., Chakraborty, S., & Schultz, S. R. (2016). Visual Receptive Field Properties of Neurons in the Mouse Lateral Geniculate Nucleus. *PLOS ONE*, *11*(1), e0146017. https://doi.org/10.1371/journal.pone.0146017

Taube, J. S. (1995). Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *15*(1 Pt 1), 70–86. https://doi.org/10.1523/JNEUROSCI.15-01-00070.1995

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *10*(2), 436–447. https://doi.org/10.1523/JNEUROSCI.10-02-00436.1990

Thomas. (2018). *Summa Theologica Complete in a Single Volume*. Coyote Canyon Press.

Thornberry, C., Cimadevilla, J. M., & Commins, S. (2021). Virtual Morris water maze: Opportunities and challenges. *Reviews in the Neurosciences*, *32*(8), 887–903. https://doi.org/10.1515/revneuro-2020-0149

Thuen, M., Singstad, T. E., Pedersen, T. B., Haraldseth, O., Berry, M., Sandvig, A., & Brekken, C. (2005). Manganese-enhanced MRI of the optic visual pathway and optic nerve injury in adult rats. *Journal of Magnetic Resonance Imaging*, *22*(4), 492–500. https://doi.org/10.1002/jmri.20400

Tinbergen, N. (1964). *Social behavior in animals, with special reference to vertebrates* (2d ed). Methuen.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Ul'ianetskaia, P. O. (1972). [Pathology of idiopathic myocarditis]. *Kardiologiia*, *12*(5), 61–65.

Van den Bergh, G., Zhang, B., Arckens, L., & Chino, Y. M. (2010). Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys. *The Journal of Comparative Neurology*, *518*(11), 2051–2070. https://doi.org/10.1002/cne.22321

Verma, I. C., & Singh, M. (1975). Letter: Down syndrome in India. *Lancet (London, England)*, *1*(7917), 1200. https://doi.org/10.1016/s0140-6736(75)93193-1

Vig, R. G. (1979). Successful prosthodontics for the general dentist. Third of a series. Boxing the impressions. *CAL [Magazine] Certified Akers Laboratories*, *43*(6), 9–13.

Vilarroya, O. (2017). Neural Representation. A Survey-Based Analysis of the Notion. *Frontiers in Psychology*, *8*, 1458. https://doi.org/10.3389/fpsyg.2017.01458

Villarreal-Silva, E. E., González-Navarro, A. R., Salazar-Ybarra, R. A., Quiroga-García, O., Cruz-Elizondo, M. A. D. J., García-García, A., Rodríguez-Rocha, H., Morales-Gómez, J. A., Quiroga-Garza, A., Elizondo-Omaña, R. E., De León, Á. R. M.-P., & Guzmán-López, S. (2022). Aged rats learn Morris Water maze using non-spatial search strategies evidenced by a parameter-based algorithm. *Translational Neuroscience*, *13*(1), 134–144. https://doi.org/10.1515/tnsci-2022-0221

Vincze, O., Vágási, C. I., Pap, P. L., Osváth, G., & Møller, A. P. (2015). Brain regions associated with visual cues are important for bird migration. *Biology Letters*, *11*(11), 20150678. https://doi.org/10.1098/rsbl.2015.0678

Von Saint Paul, U. (1982). Do Geese Use Path Integration for Walking Home? In F. Papi & H. G. Wallraff (Eds.), *Avian Navigation* (pp. 298–307). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-68616-0_30

Vorhees, C. V., & Williams, M. T. (2006). Morris water maze: Procedures for assessing spatial and related forms of learning and memory. *Nature Protocols*, *1*(2), 848–858. https://doi.org/10.1038/nprot.2006.116

Walton, K. L. (1993). *Mimesis as make-believe: On the foundations of the representational arts* (1. paperback ed). Harvard University Press.

Wang, R., & Spelke, E. (2002). Human spatial representation: Insights from animals. *Trends in Cognitive Sciences*, *6*(9), 376. https://doi.org/10.1016/s1364-6613(02)01961-7

Warner, T. A., Stafford, N. P., Rompala, G. R., Van Hoogenstyn, A. J., Elgert, E., & Drugan, R. C. (2013). Intermittent swim stress causes Morris water maze performance deficits in a massed-learning trial procedure that are exacerbated by reboxetine. *Pharmacology, Biochemistry, and Behavior*, *113*, 12–19. https://doi.org/10.1016/j.pbb.2013.09.014

Weyand, T. G. (2016). The multifunctional lateral geniculate nucleus. *Reviews in the Neurosciences*, *27*(2), 135–157. https://doi.org/10.1515/revneuro-2015-0018

Whishaw, I. Q., & Jarrard, L. E. (1996). Evidence for extrahippocampal involvement in place learning and hippocampal involvement in path integration. *Hippocampus*, *6*(5), 513–524. https://doi.org/10.1002/(SICI)1098-1063(1996)6:5<513::AID-HIPO4>3.0.CO;2-J

Whishaw, I. Q., & Tomie, J.-A. (1996). Of Mice and Mazes: Similarities Between Mice and Rats on Dry Land But Not Water Mazes. *Physiology & Behavior*, *60*(5), 1191–1197. https://doi.org/10.1016/S0031-9384(96)00176-X

Woolley, D. G., Mantini, D., Coxon, J. P., D'Hooge, R., Swinnen, S. P., & Wenderoth, N. (2015). Virtual water maze learning in human increases functional connectivity between posterior hippocampus and dorsal caudate. *Human Brain Mapping*, *36*(4), 1265–1277. https://doi.org/10.1002/hbm.22700

Wörtwein, G., Saerup, L. H., Charlottenfeld-Stapov, D., & Mogensen, J. (1995). Place Learning by Fimbria-Fornix Transected Rats in a Modified Water Mazea. *International Journal of Neuroscience*, *82*(1–2), 71–81. https://doi.org/10.3109/00207459508994291

Zaporozhan, J., Ley, S., Unterhinninghofen, R., Saito, Y., Fabel-Schulte, M., Keller, S., Szabo, G., & Kauczor, H.-U. (2006). Free-breathing three-dimensional computed tomography of the lung using prospective respiratory gating: Charge-coupled device camera and laser sensor device in an animal experiment. *Investigative Radiology*, *41*(5), 468–475. https://doi.org/10.1097/01.rli.0000208926.98693.b6