Synthese (2010) 177:261–283
DOI 10.1007/s11229-010-9849-5

Reconsidering 'spatial memory' and the Morris water maze

Jacqueline A. Sullivan

Received: 15 June 2010 / Accepted: 31 October 2010 / Published online: 25 November 2010

Abstract The Morris water maze has been put forward in the philosophy of neuroscience as an example of an experimental arrangement that may be used to delineate the cognitive faculty of spatial memory (e.g., Craver and Darden, Theory and method in the neurosciences, University of Pittsburgh Press, Pittsburgh, 2001; Craver, Explaining the brain: Mechanisms and the mosaic unity of neuroscience, Oxford University Press, Oxford, 2007). However, in the experimental and review literature on the water maze throughout the history of its use, we encounter numerous responses to the question of "what" phenomenon it circumscribes ranging from cognitive functions (e.g., "spatial learning", "spatial navigation"), to representational changes (e.g., "cognitive map formation") to terms that appear to refer exclusively to observable changes in behavior (e.g., "water maze performance"). To date philosophical analyses of the water maze have not been directed at sorting out what phenomenon the device delineates nor the sources of the different answers to the question of what. I undertake both of these tasks in this paper. I begin with an analysis of Morris's first published research study using the water maze and demonstrate that he emerged from it with an experimental learning paradigm that at best circumscribed a discrete set of observable changes in behavior. However, it delineated neither a discrete set of representational changes nor a discrete cognitive function. I cite this in combination with a reductionist-oriented research agenda in cellular and molecular neurobiology dating back to the 1980s as two sources of the lack of consistency across the history of the experimental and review literature as to what is under study in the water maze.

J. A. Sullivan (✉)
Department of Philosophy, University of Alabama at Birmingham, HB 414A,
900 13th Street South, Birmingham, AL 35294-1260, USA
e-mail: jas1@uab.edu

Keywords Cognition · Experimental learning paradigm · Mechanisms · Reliability · Spatial memory

1 Introduction

Philosophers of science who have sought to understand the nature of neuroscientific explanation have emphasized the importance of experiments for producing the data requisite to identify the targets of such explanations. For example, Bickle (2006) views the use of "well-accepted behavioral protocols" imported into cognitive neurobiology from cognitive psychology as fundamental for this purpose (Bickle 2006). William Bechtel, referring to the detection of *mental mechanisms* in cognitive science and neuroscience claims that in these domains "often experiments are required to delineate the phenomenon for which a mechanism is responsible" (Bechtel 2008, 37). Craver and Darden (2001, 122) emphasize the importance of "the accepted experimental protocols for producing, manipulating, and detecting" phenomena in cognitive neurobiology, such as *spatial memory*. Yet, despite this general consensus that experiments are fundamental for circumscribing explanatory targets in the contemporary cognitive sciences and neurosciences, scarce effort has been directed at systematically evaluating on a *case-by-case* basis *how* experimental learning paradigms (Sullivan 2007, 2009) are used to circumscribe phenomena and *what* they actually circumscribe.

An interesting case study for addressing such questions is the Morris water maze. In 1981, nearly one decade after a set of important discoveries concerning the role of the hippocampus in mammalian learning and memory (e.g., O'Keefe and Dostrovsky 1971; Bliss and Lømo 1973),[1] Richard Morris described an apparatus that rapidly became the premier device for studying the cellular and molecular mechanisms of spatial learning and memory in the rodent (see Brandeis et al. 1989). The *water maze* was, as of 2001, "one of the most frequently used laboratory tools in behavioral neuroscience", with over 2,000 papers citing its use from 1989 to 2001 alone (see D'Hooge and De Deyn 2001). The widespread use of the maze in the scientific literature prompted Craver and Darden that same year to appeal to it as a viable case study for grounding an understanding of "the continuing discovery of the mechanism of spatial memory" (Craver and Darden 2001, 112) in cognitive neurobiology.

However, despite its popularity as an experimental paradigm in neuroscience and as a case study for thinking about mechanisms in philosophy of neuroscience, what remains unclear is precisely "what" investigators who train rodents in the water maze are actually investigating. In the context of the experimental and review literature "Morris water maze performance" and "spatial memory" are used to refer to what is under study. Within the category of cognitive functions alone, "spatial learning", "spatial memory", "spatial navigation" and "spatial cognition" are all put forward to capture it. Although various options are on offer for conceptualizing it, experiments continue to be undertaken in order to discover its cellular and molecular mechanisms. Philosophical analyses of the water maze to date, on the one hand, characterize it as if it delineates a discrete phenomenon, namely, "spatial memory"

[1] I describe these discoveries at the beginning of Sect. 4, below.

(e.g., Craver and Darden 2001; Craver 2007). However, the New Mechanists also allow for the possibility that explanandum phenomena, (i.e., the referent(s) of "spatial memory") are not required to be discrete because in the process of mechanism discovery, they are subject to change. Although there is truth to this claim, it has only served to shift the focus of philosophical analyses away from what makes experimental paradigms like the water maze interesting case studies for thinking about the kinds of problems that may arise for generating mechanistic explanations when such paradigms fail to delineate what investigators claim they do and for understanding why in some cases multiple terms are put forward in order to capture "what" is under study.[2]

As an initial step towards understanding what exactly is produced, detected and measured in the water maze, the primary target of my analysis in this paper is Richard Morris's original study (Morris 1981) using the device. I begin in Sect. 2 by briefly describing how Craver and Darden (2001, and e.g., 2007) characterize the discovery of the mechanism of spatial memory in experiments using the water maze, and point out that they use the term "spatial memory" to refer to several different kinds of effects that the maze produces. In Sect. 3, I appeal to the theoretical literature in cognitive neurobiology as evidence that investigators differ in opinion with respect to which kind of effect they intend to delineate in such experiments, and I identify problems that arise with respect to the delineation of each one. This essentially constitutes a framework that I use in Sect. 4 as a basis for determining what Morris succeeded and failed to circumscribe in his original study using the water maze. I demonstrate that he emerged from this study with an experimental learning paradigm that at best circumscribed a discrete set of observable changes in behavior. However, it did not delineate either a discrete set of representational changes or a discrete cognitive function. I cite this in combination with a reductionist-oriented research agenda in cellular and molecular neurobiology dating back to the 1960s as two sources of the lack of consistency across the history of the experimental and review literature as to what is under study in the water maze.

2 Mechanisms and the Morris water maze

In their analysis of the Morris water maze as a primary instrument involved in "the continuing discovery of the mechanism of spatial memory" (2001, 112), Craver and Darden indirectly reveal an interesting problem in need of philosophical analysis. In

[2] My interest in this issue is driven by a concern that mechanistic explanations have been characterized as one form of reductive explanation (e.g., Bechtel 2009). If philosophical analyses of experimental paradigms in cognitive neurobiology fail to identify what such paradigms circumscribe, then they may take the scientists at their word that what is being circumscribed are cognitive functions (although not all cognitive neurobiologists would claim this. See Sect. 3 below), when the paradigms at best circumscribe observable changes in behavior. Yet, if the ultimate targets of mechanistic explanations in cognitive neurobiology are nothing over and above observable changes in behavior, then there is no sense in which such explanations are reductive. They are instead eliminative explanations in which a functional/representationalist ontology for identifying psychological concepts is traded in for a behaviorist one (see e.g., Machamer 2009; Sullivan 2009, 518, fn. 3).

order to characterize the nature of the problem, it is relevant to briefly introduce the "standard version" of the water maze.[3]

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. The standard water maze experiment consists of two training conditions and at least one probe trial. In the variable placement fixed visible platform condition (from herein "the visible condition"), a visible platform that protrudes slightly above the water's surface is placed into one of the four quadrants of the pool. While its location remains fixed across training trials, the placement of the rat in the pool varies randomly with respect to the four cardinal positions (i.e., N, S, E, W). The variable placement fixed hidden platform condition (from herein "the hidden condition") is identical in structure to the visible condition, except that the platform is silvery-white and hidden just beneath the water's surface so as to be undetectable to a rodent in the maze. On each training trial, the swim path of an animal in the maze, the length and direction of the angle of that path, and the time it takes it to find the platform ("escape latency") are measured. After a series of training trials (typically no less than 15, often at least 20), rodent subjects are placed into a platform-less pool. During a classic probe trial, the amount of time that the animal spends in the four quadrants of the maze as well as the number of times it crosses in and out of that quadrant of the pool in which the platform was located during training are measured.[4]

Craver and Darden (2001) identify several possibilities with respect to what is under study in the hidden condition of the water maze, in which it is supposed that a rat learns the location of the hidden platform exclusively by appeal to the distal room cues[5] (Craver and Darden 2001, 122). First, they suggest that it is the capacity or faculty of spatial memory, which they define as "roughly, the ability to learn to navigate through a novel environment" (Craver and Darden 2001, 112). However, they make other claims that indicate that the phenomena under study may include processes such as the "acquisition, storage and retrieval of spatial memories" (Craver and Darden 2001, 114). Then, in their representation of the "levels in the hierarchical organization of spatial memory" at the top-most level the phenomenon identified is "mouse navigating Morris water maze" (Craver and Darden 2001, 118, Figure 6.4), which is a different kind of activity than learning how to navigate and is not identical to the acquisition, storage and retrieval of spatial memories, although all of these activities may indeed be prerequisite for navigation. Craver and Darden classify all of these phenomena as "spatial memory". This suggests that contrary to what they claim, there are multiple different explanatory targets that are captured by this designation than

[3] When I say "standard version", I mean something very basic about the structural features of the water maze, which may be separated from the variety of "sub-protocols" that are used in conjunction with these features (Sullivan 2009).

[4] It is common in the review literature for the "standard version" to include only the hidden condition. I am presenting both conditions here for the sake of simplifying my analysis in Sect. 3.

[5] I will revisit this supposition in Sect. 4.

simply "the ability to learn to navigate through a novel environment" or "navigating the water maze".

Craver and Darden also describe spatial memory as "involv[ing] the formation of an internal spatial representation—a cognitive map—by which different locations and directions in the environment can be assessed" (2001, 122). In their hierarchical depiction of the mechanism of spatial memory, however, the formation of the spatial representation, which some may regard as the spatial memory itself, is situated at a different level than the mouse's navigation behavior. Yet on the basis of the figure, the navigation behavior appears to be the target of the mechanistic explanation rather than the spatial memory or its formation. In his recent book, *Explaining the brain: Mechanisms and the mosaic unity of neuroscience* (2007, 166, Figure 5.I), Craver offers a similar analysis of the water maze, spatial memory and its mechanisms as that offered in the 2001 paper. While he admits that "there are many different spatial memory phenomena", he leaves the question of what phenomena the water maze circumscribes open (Craver 2007, 165, fn. 4).

I think the ambiguity in Craver and Darden's treatment of the phenomenon that the water maze delineates stems from at least three sources. The first is a commitment to Bogen and Woodward's (1988) claim that "data are the evidence for phenomena" (Craver and Darden 2001, 122). Craver and Darden indicate that experimental arrangements like the Morris water maze are used to produce data that serve as evidence for a phenomenon like spatial memory. However, they do not explicitly acknowledge that the data that such experimental arrangements yield may be used to detect multiple distinct kinds of phenomena (see for example, Feest, forthcoming; Sullivan 2009, forthcoming). Second, their analysis directly reflects the lack of clarity in the experimental and review literature in cognitive neurobiology with respect to what the water maze circumscribes. Third, I take it that Craver and Darden's aim (2001; Craver 2007) is exclusively to understand the hierarchical structure of mechanistic explanations and the spatial and temporal constraints characteristic of the mechanisms themselves, which are, after all, on the side of the explanans rather than the explandum. Furthermore, the "New Mechanists" including Craver and Darden (2001; Craver 2007) and Bechtel (e.g., 2008) allow for the possibility that mechanism discovery may proceed without discrete explanatory targets.

Craver and Darden's analysis of spatial memory and the water maze in combination with what I have characterized above as a lack of clarity in the scientific literature as to what the maze delineates prompts two questions. First, what phenomenon does the maze actually circumscribe? My strategy for addressing this question is to go back to Morris's original study (1981) using the water maze, prior to its rapid and widespread implementation in cognitive neurobiology in 1980s and 1990s. Second, why are so many options on offer in the experimental and review literature with respect to what it circumscribes? As I will demonstrate, answering the first question by way of Morris's original study will shed light on the answer to the second question. In order to structure my analysis of the case study, I offer a framework for thinking about the kinds of phenomena that experimental learning paradigms like the water maze circumscribe in Sect. 3.

3 What do experimental learning paradigms circumscribe?

As I understand the term, an experimental learning paradigm is a standard set of procedures for producing, measuring and detecting a form of learning and memory in the laboratory (Sullivan 2007, 2009). It specifies *how* to produce a form of learning or memory, which includes a description of what stimuli are to be presented to an organism, what the spatial and temporal arrangement of those stimuli ought to be, and when and how many times they should be presented to the organism during pre-training, training and post-training or testing. It also identifies the response variables to be measured during pre-training, training, and post-training/testing and how to measure them using equipment that is designed for this purpose. Finally, it specifies how to detect the form of learning or memory when it occurs, by identifying what the comparative measurements of the selected response variables have to equal in order to ascribe to the organism that form of learning or memory the investigator intends to produce. This is essentially an operational definition (see for example, Chang 2004, 2009; Feest 2003, 2005), which I take to be built directly into the design of an experimental learning paradigm.

In cognitive neurobiological experiments in which learning paradigms are used, the response variables that are measured are always behaviors. Examples of behavioral responses include freezing, sniffing, and food consumption. A baseline response to a stimulus or set of stimuli measured before training in a paradigm typically constitutes one data point. The behavioral response to that same or a different stimulus arrangement after training constitutes another. A change in behavioral response may be detected by comparing these two data points. If the relevant kind of change is detected (i.e., that change specified in the operational definition), then the change in behavioral response may be classified as an instance of the corresponding type of learning or memory (e.g., conditioned fear). Although this may give the impression that the phenomenon that is being detected or that is the target of an ensuing explanation is a cognitive process or function that is distinct from the change in behavioral response, many investigators in cognitive neurobiology only claim to be detecting the latter. For example, in his textbook, *Mechanisms of Memory* (2009), the cognitive neurobiologist David Sweatt indicates this very clearly in claiming that learning is:

> the acquisition of an altered behavioral response, due to an environmental stimulus; in other words, learning is when an animal changes its behavior pattern in response to an experience. […] Note that what is defined is a change in a behavior from a pre-existing baseline. Don't get confused that I am defining learning as a response to an environmental stimulus, but rather as an alteration in that response due to an environmental stimulus. An animal has a baseline response, experiences an environmental signal, and then has an altered response different from its previous response (Sweatt 2009, 3).[6]

[6] It is worthwhile to note that the water maze contains multiple different kinds of environmental stimuli, making the tracking of behavioral responses to stimuli exceedingly difficult. Furthermore, investigators have historically been unclear with respect to how to itemize these stimuli. For example, the distal room cues may be treated as independent or compound stimuli (and compounds involving different groupings of stimuli).

To take an example that illustrates Sweatt's point, in a fear-conditioning paradigm, the level of freezing that an organism exhibits to the tone before training (i.e., baseline response to tone) may be compared to the level of freezing the investigator measures after the tone has been paired with a high-intensity shock. By appealing to these two data points, an investigator is able to detect a change in the response variable—i.e., a change in the organism's behavior. If a significant increase in freezing to tone is what is required for conditioned fear learning ascriptions, it may simply be that such ascriptions refer exclusively to the alterations in behavior, and what the paradigm is taken to circumscribe from the standpoint of the investigator is an organism-level phenomenon.

A primary problem with what John Bickle has referred to as this "reduction-in-practice" strategy in cognitive neurobiology is that investigators use the same term to classify the changes in response variables that they use to classify the realization or instantiation of a cognitive capacity or function. For example, with respect to the data obtained in a fear-conditioning paradigm, they do not refer to what they detect or what they are trying to identify the cellular and molecular mechanisms of as increases in levels of freezing in response to tone following multiple tone-shock pairings. Rather, they refer to it as "fear-conditioning" or "learned fear". This is misleading if it is exclusively the changes in behavioral response variables that are the targets of their cellular and molecular explanations and what they aim to intervene in the production of. Of course, part of the problem is, as Bickle (2006) indicates, that the vast majority of learning paradigms that cognitive neurobiologists use for the purpose of circumscribing changes in behavior originate in areas of science that *are* interested in "what" organisms learn, and those investigators do take the paradigms to circumscribe discrete cognitive capacities (e.g., social recognition). As a consequence, the paradigms are referred to in ways that suggest that they produce and detect discrete cognitive capacities. Investigators in cognitive neurobiology are thus appealing to the same terms even though they are, with few exceptions (like those I identify below), exclusively interested in behavioral changes.

However, some investigators do indicate that the data used to detect changes in behavioral response variables may also be used to detect events causally prior to those changes that are distinct from alterations in cellular and molecular activity. One suggestion is that the data are used as a basis for ascribing changes in internal representations (e.g., the formation of a tone-shock association) to an organism trained in the learning paradigm (See for example Sullivan, forthcoming). For example, in the fear-conditioning example, the baseline response to tone may be taken to indicate an absence of a tone-shock association (i.e., representation) and the post-training response to tone may be taken to indicate the presence of such a representation as a result of the experimental manipulation (i.e., training in the learning paradigm). In support of this possibility, in his textbook, *The Neurobiology of Learning and Memory: Concepts, Findings and Trends*, Yadin Dudai defines learning as "an experience-dependent generation of enduring internal representations, and/or an experience-dependent lasting modification in such representations" (1989, 6). The learning paradigm itself may be designed so as provide the organism with an experience capable of generating a representation of a specific type (e.g., tone-shock association). However, as Dudai acknowledges if only indirectly, the investigator assumes that the control exerted over

"what" the organism learns is built right into the overall experimental design and production procedures of a learning paradigm. In other words, such production procedures are supposed to produce the formation of one type of representation rather than another (or many others). The fundamental problem, however, is that even if the investigator may use those production procedures to detect the changes in behavioral response variables that he takes to be indicative of a specific type of change in internal representations, the changes in behavior reveal little as to the content of the representational changes, i.e., what is learned. In other words, the paradigm may lack the control requisite for constraining what the organism learns, and thus lack the reliability requisite for producing data that may be used to discriminate among competing claims about 'what' the organism learned or what the changes in internal representations actually are.

One approach to determining if a paradigm has the requisite kind of control is to evaluate the relationship between the production procedures, the claims an investigator makes about how an organism behaves in the context of training and testing in the paradigm, and what the investigator claims about "what" is learned on the basis of the data. However, this approach requires a detailed consideration of the features of the learning paradigm and the possibilities that it may afford or exclude with respect to what is learned. It is highly unlikely that an investigator will successfully itemize all of the possibilities with respect to what an organism may learn when run through an experimental learning paradigm. This will impact the reliability of the data production process in so far as the data may wrongly be used to discriminate claims about one type of representational change when there are in fact others (even many others). Additionally, if the representational changes are more extensive than the data are taken to indicate, yet it turns out that these changes are grouped under that same cognitive capacity or function as are the changes in behavior (e.g., if the change is in spatial representations, then there is only spatial memory), many discrete types of representational changes will be put forward as one (see Craver 2007 on "lumping errors"). Thus what is circumscribed will seem unitary when it is not, and similarly the target of any ensuing explanation will be treated as unitary when it is not. Such problems may go unnoticed in those instances in which the discreteness of a set of behavioral changes is used as a basis upon which to infer the discreteness of a set of representational ones.

In the experimental and review literature in contemporary cognitive neurobiology explicit appeals to changes in internal representations or "what" organisms learn are rare although they sometimes occur (e.g., Machamer 2009; Sullivan et al. 2004, 2005; Sullivan, forthcoming). It is more common in cognitive neurobiology that a learning paradigm is used to produce data in order to detect "the behavioral expression" (i.e., changes in behavioral response variables) of a cognitive capacity. Thus, if the relevant change in response variables occurs (e.g., increased levels of fear), an investigator will infer that the cognitive capacity has occurred (e.g., associative learning).[7] Similarly, in the context of intervention experiments, if inhibiting a protein kinase is shown to disrupt the behavioral expression of the capacity, then the investigator will likely assume

[7] As Morris (1989) claims, "learning and memory are fundamentally psychological concepts inferred from alterations in behavior in response to experience" (3052).

that she has disrupted the cognitive function itself or some stage of that function (e.g., memory encoding or retrieval).

One problem with inferring cognitive capacities from data produced by means of experimental learning paradigms is that those capacities are pitched at a fairly course-grain (e.g., "spatial memory", "social recognition memory")—a grain that likely includes numerous cognitive capacities and implicates many brain systems. Thus, it is often unclear in the context of intervention experiments, which capacity is being disrupted when performance in an experimental paradigm is shown to be impaired.[8] If an investigator fails to consider the variety of capacities, functions, and processes that may be involved prior to, during, and after training (including during inter-trial and inter-training block intervals), then claims with respect to 'which' capacity is disrupted or impaired by such interventions will be tentative at best. This is likely one reason why claims in cognitive neurobiology are often qualified in so far as a brain area or molecule is said to be "implicated in" a cognitive function, which suggests that it is still an open question as to what functions or brain areas the overall cognitive function includes and what role(s) the molecule plays. However, in such cases, if an investigator makes interpretive claims on the basis of the data to the effect that she has circumscribed a discrete cognitive function and identified its mechanism, that function is likely not as discrete as the term used to refer to it implies. The ultimate target of the mechanistic explanation that the paradigm is used to generate will also not be discrete, although this may be missed given that the purported behavioral expression of the cognitive capacity and its absence in the presence of intervention experiments are taken to be discrete.

With this basic framework for thinking about the kinds of problems that may arise with respect to using experimental paradigms to circumscribe learning and memory phenomena in cognitive neurobiology, I turn to an analysis of the case study.

4 The Morris water maze

I described the basic structural features of the "standard version" of the Morris water maze in Sect. 2 above. In this section, I am primarily interested in determining what Richard Morris (a) intended to circumscribe by designing and implementing the water maze in the laboratory, (b) what he actually managed to delineate by training rats in it, and (c) what he claimed to have achieved. To this end, I revisit Morris' first published research study using the water maze and evaluate the reliability of the data production process for discriminating between competing claims about 'what' it circumscribed.

Morris begins the paper by claiming "the distinction between 'proximal' and 'distal' orientation" is "crucial to [its] design" (Morris 1981, 240). He defines 'proximal' orientation as learning to approach a goal that is detectable by one or more senses, whereas he understands 'distal orientation' to imply "learning about the spatial location of a goal relative to distal cues" (Morris 1981, 240). As Morris indicates, at the time of his writing the distinction was purely a theoretical one. "Proximal orientation", if understood to be a form of learning, was taken to involve mechanisms of

[8] Of course, this corresponds to the claim I made that the organism may be learning many things in the context of a learning paradigm.

associative learning. "Distal orientation", in contrast, was thought to involve 'place learning'—learning the spatial location of a target relative to distal cues. Morris admits that "it [was] not yet clear whether [the distinction] represent[ed] a valuable [one] with respect to the underlying mechanisms of orientation" (Morris 1981, 240). However, an encouraging set of discoveries in neurophysiology in the early 1970s led many investigators to believe that it might be, which prompted Morris to aim to develop a paradigm that could dissociate the two. It is worthwhile to briefly put the development of the water maze into this broader historical context, to pinpoint the theoretical ideas that shaped its development.

First by 1973, the literature in experimental psychology was rich with examples of hungry rats demonstrating improved performance across training trials in locating food rewards hidden in mazes. During training, their paths to the goals became more direct and they took less time to reach them. Stimulus–response theorists (S–R), stimulus–stimulus (S–S) theorists and field theorists put forward competing hypotheses to explain such changes in behavior (e.g., see Tolman 1948). S–R theorists proposed that rats learned how to navigate via the association of local cues (e.g., a corner in the maze) with motor responses (e.g., a left or right turn) [i.e., "proximal orientation"]. S–S theorists assumed that associations were instead forged between discrete sets of local cues (e.g., an odor and a corner of the maze) [i.e., "proximal orientation"]. In contrast, proponents of field theory claimed that across trials a "cognitivelike map of the environment gets established in the rat's brain" (Tolman 1948), as the rat attends exclusively to the spatial relations between the goal and the extramaze cues (i.e., "distal orientation"/"place learning") (Tolman et al. 1946, 1947; Restle 1957).

All three competing hypotheses accommodated the data equally well.[9] However, two physiological findings in the early 1970s were regarded as tipping the evidential balance in favor of cognitive map theory. First, O'Keefe and Dostrovsky (1971) discovered what they dubbed "place cells" in the hippocampus—a set of pyramidal cells that may fire bursts of action potentials in response to specific spatial locations when a rat freely moves about in an environment.[10] This discovery lent support to the idea that as rats explore a maze, thus sampling that environmental space, coordinates in that space come to be represented in the hippocampus, such that no location prompt ("S") would be requisite to elicit a turn ("R"). The rats could instead *distally orient* to a target by appeal to the representation of the spatial relations holding between it and the extramaze cues that was distributed across hippocampal place cells.

The second finding was Terge Lømo and Tim Bliss's discovery of a long-lasting potentiation (LLP, later known as long-term potentiation (LTP)) in the hippocampus of the anesthetized rabbit, which instantiated the very features that Hebb (1949) had ascribed to the neurophysiological correlate of learning and memory. LTP offered a plausible neurophysiological mechanism for the formation of cognitive maps distributed across place cells in the hippocampus (O'Keefe and Nadel 1978).[11] This finding

[9] There were additional hypotheses that also accommodated the data (see Tolman 1948).

[10] "Place cells" may fire in response to non-place cues as well (see for example, Redish 2001).

[11] It should be pointed out, however, that persistent increases in synaptic strength could also be regarded as the mechanism for the formation of S–S and S–R connections in an organism's brain. In fact, Hebb (1949) intended his postulate to capture how such associative connections were forged between neurons.

was compatible with the idea that once a representation was stored across place cells in the hippocampus, a rodent could appeal exclusively to it in order to find its way in a maze.

The only item that was missing, then, was an experimental learning paradigm that would serve to connect the neurophysiological findings to learning and memory. Morris claimed that the reason that none of the available mazes could serve this function is that they failed to solve what he dubbed "the local cue problem" (Morris 2003, 244). All the available mazes contained local cues such that it was impossible to determine whether rats oriented to a goal in a maze via (1) local or 'proximal' cues, (2) a combination of 'local' cues and motor responses (3) the extra-maze 'distal' cues" (e.g., cues outside of the maze apparatus, including, e.g., radiators, bookcases, doors of the laboratory room) or (4) all of the above. Local cues (e.g., corners in mazes, odor trails) were thus confounds that would have to be eliminated in any task that sought to circumscribe distal orientation or place learning. Morris's solution to the "local cue problem" was the hidden condition of the water maze.[12]

Morris's study may thus be understood to begin with the following empirical question: Can rats "rapidly learn to locate an object that they can never see, hear, or smell provided it remains in a fixed spatial location relative to distal room cues (Morris 1981, 239)?"[13] This empirical question corresponds to two competing hypotheses: (1) rats *can* rapidly learn to locate a hidden object provided it remains in a fixed spatial location relative to distal room cues and (2) rats *cannot* do so. At first blush, these competing hypotheses may seem modest and non-committal with respect to what rats learn and how—as if Morris is interested in producing data requisite to relegate between competing claims about something closer to observable changes in behavior than changes in internal representations or cognitive functions. However, he intends the data to establish that learning in the hidden condition involves the formation of a certain kind of representation (i.e., a cognitive map) and the activities of a discrete type of cognitive function (i.e., place learning (which already includes distal orientation)). Evidence for this point is that the experimental design and protocols that

[12] Morris, however was undertaking a project that some investigators had argued could not get off the ground. For example, in a paper in 1957, the experimental psychologist Frank Restle, in a paper that Morris (1981) cites, indicated in response to the work of psychologists like Tolman, Ritchie and Kalish that "place learning" was not a valid construct:

> there is nothing in the nature of a rat which makes it a 'place learner', or a 'response' learner. A rat in a maze will use all relevant cues, and the importance of any class of cues depends on the amount of relevant stimulation provided as well as the sensory capacities of the animal. […]. The writer's general conclusion is that further 'definitive' studies of the place-vs.-response controversy, to prove that rats are by nature either place or response learners would be fruitless since the issue is incorrectly drawn" (Restle 1957, 227).

[13] As Morris claims:

> One day, it occurred to me that rats might be able to learn while swimming and that this might help solve the local cue problem. I wondered if they could escape from water onto a platform that was hidden beneath the water surface and so was neither visible, audible, offered no olfactory cues and could not be identified using somatosensory cues until the animal had already successfully navigated to it. (Morris 2003, 644)

correspond to the water maze are intended to constrain how and what rats trained in each condition of the maze learn. But was the data production process *reliable* in so far as it offered such constraints? In order to answer this question, I will focus on those data obtained from training rats in the hidden condition as well as those additional data that serve to strengthen the claims that Morris makes about the data obtained in the hidden condition.

In Morris's original study the water maze was placed in the center of a room containing four fixed visual distal room cues, one per wall (shelves, a door, a window, and a cupboard) (Morris 1981, 242). Rats were given two pre-training trials (180 s/day for 2 days) to acclimate them to swimming in the pool. They were then divided into four groups (eight per group). Each group was run through 20 training trials ("escape trials") of one of four maze conditions spread across 3 days. In addition to the visible and hidden platform conditions (described in Sect. 2), one group of rats was run through a random visible condition, in which the placement of the rat in the pool as well as placement of the visible platform in the pool randomly shifted across trials. A fourth group was trained with the hidden platform randomly shifting from quadrant to quadrant across all 20 trials.[14]

Morris used an electronic timer to measure each rat subject's latency to reach the platform on each trial (Morris 1981, 242). He then averaged the performance of subjects in each group on each trial (e.g., Morris 1981, 246, Figure 3). A video camera positioned above the center of the pool was used to record each animal's trajectory through the pool. These recordings were used to determine the length and angle of an animal's swim path from the point of placement in the maze to its arrival atop the platform on the last four training trials (Morris 1981, 241). Morris also calculated the mean path length and median direction of the path for each group with respect to these last four training trials (Morris 1981, 247). On the testing or "probe trials", which immediately followed training, he first placed rats into a platform-less pool, and then into a pool with a platform in either a different or a fixed location. During these probe trials he calculated the amount of time that rats spent in each of the four quadrants of the pool and how many times they crossed in and out of each of the four quadrants.

Morris defined "spatial localization" generally as the ability to learn to locate a platform, irrespective of whether it is visible or hidden, in the pool. He operationally defined it as a significant decrease in escape latency across training trials. He determined such decreases by comparing the data points obtained for each group across all 20 training trials. The other measurements he took were used to characterize the qualitative features of the observable changes in behavior across groups. For example, the angle of the path in which the animal headed on the last four training trials was taken to indicate the accuracy of its directional heading towards the platform. He defined "spatial bias", a feature of the pattern of a rat's swimming path in the pool during probe trials, as more time spent in that quadrant of the pool in which the platform had been located during training than in any other quadrant of the pool.

Morris engages in a comparative analysis of the data obtained across all four conditions in order to discriminate between the two competing hypotheses I identified

[14] Morris dubbed these four groups: (1) "Group Cue + Place", (2) "Group Place", (3) "Group Cue Random", and (4) "Group Place Random".

above. First, taking the data obtained from the hidden condition in isolation, he claims that rats trained in the hidden condition exhibited rapid and significant decreases in their escape latencies across trials that were comparable to those of rats trained in both visible platform conditions. Secondly, he points out that on the last four training trials, the swim paths that they took to the platform were as direct as those taken by rats trained in the visible platform conditions. He also notes that compared to rats trained in the visible platform conditions, the "spatial bias" of rats trained in the hidden condition is more pronounced. He appeals to the data obtained from the random hidden condition as a basis for claiming that the spatial relationships between the platform and distal room cues being kept fixed were essential for the rapid decreases in escape latency exhibited by rats trained in the hidden condition. He also uses these data as a basis for making claims about 'what' rats in two hidden groups learned. He suggests that rats in the random hidden condition must learn only that "escape is possible" (Morris 1981, 246), which accounts for the slight decreases in their escape latency as well as the fact that they never reach the levels of performance of rats trained in the hidden condition.[15]

Although this one study may be regarded as insufficient for establishing the *reliability* of the hidden condition of the water maze for detecting a discrete set of changes in behavioral response variables, the data that Morris presents are sufficient to establish that the swimming behavior of eight rats changes significantly when they are trained in the hidden condition of the water maze and that they exhibit a set of behaviors post-training and during testing that were not exhibited prior to such training. One way to describe the changes is simply that rats trained in the hidden condition exhibit improved performance in several dimensions (i.e., decreases in escape latency, more accurate heading to the platform, a bias for the training quadrant)—i.e., their behavior changes. Morris in fact begins by providing a modest interpretation of the data in so far as he suggests they may be taken to indicate that rats trained in the hidden condition are successful with respect to "localization of a hidden object" (Morris 1981, 252). However, the paper does not end with this claim. It is followed by a question, namely "How did the rats do this?" (Morris 1981, 252).

I interpret the very raising of this question as an admission on Morris's part that the data could not be used to discriminate between the two competing hypotheses I identified above if they are taken to include claims about what and how rats trained in the hidden condition learned. In raising it Morris may be regarded as admitting that the data production process was reliable in so far as it produced a set of data that indicated that rats trained in the fixed hidden condition *located the platform* time and time again. However, accounting for the decreases in escape latency across trials observed in this group (detected by a comparison across individual data points for each training trial run) was a different matter. The decreases indicated that the rats got better at finding the platform over time. Yet, those decreases could not be used to detect "what" the representational changes were, i.e., "what" rats trained in the hidden condition learned, or "how" they learned it (i.e., what the function was).

[15] I will return to this point in Sect. 5.

The interesting question, then, if the data produced in the hidden condition of the water maze did not discriminate between the two competing hypotheses—if it failed to reveal how and what rats trained in the hidden condition learn, on what basis does Morris go on to infer that the data may be used to serve both of these functions?[16]

Towards the end of the paper, Morris draws a distinction between (a) the nature of the change in the rats' representation of the spatial environment across training trials (i.e., representational changes) and (b) the mechanisms productive of those representational changes (i.e., the capacity—associative mechanisms versus place learning mechanisms). He appeals to the "behavioral flexibility" of rats trained on the hidden condition of the platform as the basis upon which to infer that the "form of the stored representation" is most likely a cognitive map. He identifies two "separate demonstrations of this putative flexibility", namely, that these rats rapidly learn the location of the hidden platform and during probe trials they exhibit "a novel search strategy" in so far as they search for the platform in its previous location in the pool. Nothing in the data provides him with the precise content of "what" rats in the hidden condition learn. He tries to get at the nature of that content indirectly based on received views about the kinds of behaviors that cognitive representations in contrast to associative representations afford.

With respect to whether or not Morris thinks that the hidden condition of the water maze circumscribes a discrete learning capacity to go along with the purported representational changes—I think we encounter an ambiguous answer. Given his endorse-

[16] At this point in his paper, Morris identifies three competing hypotheses that respond to the question of how rats trained in the hidden condition learn the location of the platform as well as what they learn. His very positing of these hypotheses suggests that his attempt to constrain the range of competing hypotheses as to how rats in the hidden condition learn the location of the platform may have failed, so he requires additional data to discriminate among these competing hypotheses. Interestingly, these hypotheses correspond directly to that set of competing hypotheses (i.e., S–S, S–R and field theories) that other maze experiments could not be used to produce data to discriminate between because they all contained "local cues". The first hypothesis is that rats trained in the hidden condition relied exclusively on the relationship between the hidden platform and the extra-maze cues in order to locate it (i.e., "place learning" or cognitive map theory). The second posits that "rats learned four distinct cue-approach responses" in relationship to the extra-maze cues (S–R theory) (Morris 1981, 253). The third suggests that the rats "learned to swim off at various angles from the side walls from the different starting positions" (253) and thus locate the platform (S–S theory).

To discriminate among these hypotheses, Morris conducted an additional experiment, namely "a transfer test". His rationale for running this experiments was to test the predictions of these three hypotheses under conditions in which rats were repeatedly trained in one condition of the maze, but then tested on another. Whereas the associative learning theories both predict that rats trained in one condition of the maze will not be able to generalize to another, the cognitive map theory predicts that they will due to the behavioral flexibility that the representations formed by such learning affords. So, Morris trained a group of rats in the water maze for a total of 15 escape trials. All three groups were placed into the pool at the same starting location (W) with the location of the hidden platform (NE) being held constant across trials. By the end of the training trials, all three groups exhibited significant decreases in escape latency. They were then divided into three groups and each group was exposed to a different testing condition. The first group was tested on a condition in which the angular relationship between the starting location of the rat in the pool and the location of the platform was identical to that angular relationship on which they had been trained (W–NE). The angular configurations on which they were tested thus were N–SE, E–SW, and S–NW. In the second group the platform remained in the same location, but the rats were placed into the pool at one of the three other starting locations (i.e., N, S, E). In the third group, the conditions remained the same as during training.

ment of cognitive map theory and his referring to rats in the hidden condition as "Group Place", we might anticipate that he would claim that these rats orient distally and are place learners. However, at the end of the paper he claims that "the results provide support for the cognitive mapping theory of spatial localization but no definite evidence that the processes underlying the formation of a map or its use in behavior are distinct from those processes explored in traditional studies of associative learning" (Morris 1981, 259). This suggests that the nature of the cognitive function is for him an open question.[17] This means that coming out of the 1981 paper, Morris had put forward detection techniques that did serve to circumscribe a set of observable changes in behavior. However, there was no concept to which such changes in behavior attached with respect to an operational definition. In other words, Morris developed a means to detect a set of behavioral changes, but he had not identified what psychological function or set of representational changes they were indicative of.

As I aim to show in the next section, this state of affairs in combination with a reductionist-oriented research agenda in cellular and molecular neuroscience provided fertile grounds for leaving these questions unanswered and for the changes in behavior observed in the hidden condition to become subject to multiple different

Footnote 16 continued

Morris compared the mean escape latency for each group on the last 3 training trials to their escape latency on the first three testing trials. The mean escape latency of subjects tested on the condition in which the angular relationships were held constant increased significantly during the testing trials and on average they spent more time in that quadrant of the pool in which the platform had been located during training. In contrast, the escape latency of subjects tested on the other two conditions remained the same across the training and testing trials and the data indicated that their swim paths to the platform were both short and direct. Morris suggests that if the rats trained in the fixed angle configuration had learned to "swim off at a particular angle from the side walls" in order to locate the platform (Morris 1981, 255), they would have had no difficulty finding the platform during the testing trials. However, the increases in escape latency and the amount of time they spent in the quadrant of the pool in which the platform was located during training suggest that they learned "the place" of the platform relative to the distal room cues, rather than learning to swim at an angle from the starting position in the pool. Second, Morris claims that because subjects tested in the condition in which the location of the hidden platform was the same as that during the training trials do not exhibit an increase in escape latency when they are placed into the pool at different start locations, "they [have not] acquired any simple S–S association of swimming towards a specific distal cue" (Morris 1981, 255). He briefly attempts to turn back objections to his interpretation of these data, but then concludes that these additional data are consistent with "the cognitive mapping account" (O'Keefe and Nadel 1978) being "the most parsimonious interpretation of the data with its explicit claim that the stored representation of the distal room cues permits the generation of novel directional behavior" (Morris 1981, 227).

This second set of experiments adds little if anything to our understanding of what rats trained in the hidden condition in the first set of experiments learn. The combined data from the "fixed platform" training conditions (and probe trials for those groups) make it look as if when the spatial arrangement of the platform and distal cues is kept constant rodents will appeal to those cues in order to locate the platform. This is one interpretation of the data—merely appealing to the data from rats trained in the hidden condition to arrive at claims about their behavior does not reveal how and what those rats learn. In addition, with respect to this second set of experiments, it is not clear that Morris had exhausted all of the options for what might count as an instance of S–S or S–R learning in the water maze (e.g., see Tolman 1948) or, more specifically, he had not exhausted all of the possible options for what might constitute a response (i.e., there are other motor movements that the rats are making besides swimming at angles, for example) or a local cue for the rat.

[17] He acknowledges that it remains an open question at the end of a 1982 experimental paper as well as in a 1984 methodology paper. In fact the problem is not explicitly addressed until a 1990 paper, in collaboration with Eichenbaum and Stewart, described in Sect. 5.

classifications. For the sake of brevity, at best I offer only part of a complex story—a subset of additional factors that have contributed to oscillations in the experimental and review literature as to what the water maze delineates.

5 The Morris water maze post-1981

Neuroscientific investigations using the water maze could have gone in at least two general directions subsequent to the publication of Morris's original study. First, they could have been directed at determining the component informational and representational processes involved in rodent performance on the hidden condition of the water maze. However, they could have veered in another direction, namely, to determine the cellular and molecular mechanisms of the observable changes in behavior exhibited by rats trained in the hidden condition. There is a wide array of evidence to suggest that investigators, especially Morris, were initially engaged in both types of projects. However, because reductionist-oriented approaches—approaches that turned away from thinking about component informational and representational processes to focus on the synaptic and cellular-level processes of encoding, storage and retrieval—held sway in cognitive neurobiology, in that research domain the dominant strategy became as Bickle (2006) characterizes it "intervene cellular-molecularly and track behaviorally". This does not mean that all investigators in cognitive neurobiology turned a blind eye to interesting patterns of behavior exhibited by rats trained in the water maze during their production and intervention experiments.[18] In fact, in conjunction with their cellular and molecular studies it was customary to offer hypotheses as to the various kinds of things that rodents trained in the water maze learned, at what stages different representational and informational processes were operative and when different cognitive functions effectively "kicked in" (e.g., see Brandeis et al. 1989; D'Hooge and De Deyn 2001). These findings, however, took a back seat to the data linking cellular and molecular activity to the observable changes in behavior, and were often bracketed in the results and discussion sections in experimental studies and dealt with somewhat causally in the review literature. Primarily because Morris's original study left what (i.e., the representations) rats trained in the hidden condition learned and how (i.e., via which capacities or functions) open questions that no one in cognitive neurobiology actively sought to systematically address, different terms were put forward to designate the observable changes in behavior.

As evidence in support of these claims, I appeal to a very small subset of the experimental and review literature from 1981 to 2004. To begin, I focus on Morris's own studies using the water maze and then turn to insights contained in several representative review papers.

Between the years 1982 to 1992, Morris referred to the observable changes in behavior in the hidden condition variously as "place navigation" (e.g., 1982, 1989), "place learning" (e.g., 1986, 1989; Eichenbaum et al. 1990), "spatial localization" (1984)

[18] As I indicated in Sect. 2, investigators in cognitive neurobiology have different sensibilities with respect to what experimental paradigms circumscribe and what they take the targets of their explanations to be. The options are changes in behavioral response variables (e.g., Sweatt 2009), changes in internal representations (e.g., Dudai 1989), and the activation of cognitive capacities.

and "spatial learning" (e.g., 1984, 1989; Morris et al. 1990; Davis et al. 1992). In one set of intervention experiments, Morris et al. (1982) trained rats with hippocampal lesions in the hidden condition of the water maze. The data indicated that, compared to controls, the performance of hippocampal lesioned rats on the hidden condition was impaired. Morris and colleagues describe this as an impairment in place navigation. The designation "place navigation" appears to be intended to capture the fact that at least two discrete processes occur in the context of the hidden condition: learning the location of the hidden platform and navigating towards it. The lesioning studies in conjunction with the behavioral data did not afford the precision to identify which process was impaired, which may explain the designation. Of course, whether normal rats trained in the hidden condition were place learners or place navigators still lacked the requisite evidential support.

Beginning in 1986 and continuing until the late 1990's, Morris's intervention studies using the water maze were directed at determining the role played by *N*-methyl-D-aspartate (NMDA) receptors in 'what' was under study in the water maze. In the early 1980s a number of studies had yielded data in support of the idea that blockade of NMDA receptors by aminophosphonovaleric acid (APV) blocked the induction of activity-dependent LTP at hippocampal synapses, but neither its maintenance nor its expression. This prompted the question of whether NMDA receptor blockade may block learning (i.e., acquisition) but not memory retrieval. While acquisition and retrieval are stages in information processing, they can be correlated with observable changes in behavior in a way that enables an investigator to bypass questions of what is learned or what component processes precede or are involved in such acquisition (e.g., attention). In essence the categories offer one bottom-up strategy for parsing rodents' behavioral performance in the hidden condition into discrete stages while obscuring "what" and "how" they learn. So, the drive to link the mechanisms of LTP induction to 'what' was under study in the hidden condition prompted a reclassification of the observable changes in behavior from "place navigation" (which implies acquisition, storage and retrieval) to "place learning", which implies that the hidden condition may be used exclusively to study the impact of microinfusing APV in the hippocampus on learning and memory acquisition. Data from intervention studies conducted from 1986 to the late 1990s are now widely taken to establish that NMDA-receptor blockade specifically impairs learning in the water maze, that this impairment is dose-related and that the deficit is anterograde (e.g., Morris 1989; Morris et al. 1990; Davis et al. 1992).

After 1990, "place learning" was for the most part replaced by "spatial learning" to identify what is under study in the water maze. Several factors may have contributed to this shift. First, during the years after the original study, Morris indicated that learning in the hidden condition, "may involve" either "'cognitive mapping'" or "a 'snapshot' procedure for homing in on the correct location", which cast doubt on the idea that place learning was the cognitive function under study in the hidden condition and that the representational changes amounted to the formation of a cognitive map (Morris 1984, 58). In the early 1990s he engaged in both independent and collaborative research (e.g., 1990; Eichenbaum et al. 1990) as a means to determine "which procedure" rats trained in the hidden condition employ. Eichenbaum et al. (1990) introduced two forms of place learning that correspond to these two procedures and

dissociated them on the basis of differences in the behavioral performance of rats with fornix lesions (effecting hippocampal functioning) compared to normal rats on two conditions of the hidden water maze task. The first condition was the standard hidden condition. The second condition differed from it only in so far as rats were placed into the pool at the same starting position across trials. Eichenbaum and colleagues interpreted the data as lending support to Morris's (1981) original claim that normal rats trained in the variable hidden condition of the water maze, given the flexibility of their behaviors (i.e., returning to platform from novel start locations and spatial bias during probe trials), learn "a place represented in terms of positional relations among environmental cues and observer", (1990, 3531) rather than merely "a simple association between an individual set of cues and the behavior reinforced by successful escape" (1990, 3539). In contrast, rats without a functional hippocampus are forced to depend on associative cues (i.e., a snapshot procedure) to locate the hidden platform.[19]

However, to use the designation "place learning" to capture the observable changes in behavior in both conditions of the water maze, when in one condition the learning is described as associative whereas in the other it is a form of learning distinct from associative learning (i.e., place learning) was contrary to the definition of that form of learning dating back to Tolman. This may be one contributing factor in the diminished frequency of the usage of the term post 1990 to capture the observable changes in behavior in the standard hidden condition of the water maze and the increased frequency of the designation "spatial learning" (e.g., Davis et al. 1992; Morris et al. 1990) to capture it. However, the data obtained from these experiments also raised the possibility that rats trained in the standard hidden condition could use two kinds of strategies to locate the hidden platform—thus the designation "spatial learning" made more sense in so far as it included both possibilities.

Although I have evaluated only a subset of Morris and colleagues' research using the water maze, I think what becomes clear as we look across even this subset of the experimental and review literature is that uncertainty with respect to what rats trained in the hidden condition learn and how they learn it persisted while the use of the water maze in cognitive neurobiology post-1990 escalated (see for example Sutherland and Hamilton 2004; D'Hooge and De Deyn 2001). Between the years 1989 to 2001 alone over 2,000 research papers using the hidden condition of the water maze had been published. In concluding this section, I want to consider a couple of review papers that indicate further fluctuations in the experimental and review literature up until 2004 as to what the hidden condition delineated.

As I mentioned earlier in this section, a primary aim of contemporary cognitive neurobiology is to identify the cellular and molecular mechanisms of learning and memory. Immediately following publication of Morris's 1981 study, the water maze was implemented in a number of laboratories both for the purposes of running intervention experiments (i.e., lesioning and pharmacology) and testing out a wide array of experimental protocols used in conjunction with the maze (see Brandeis et al. 1989). The changes in behavior that Morris had produced in normal rats in the hidden condition proved to be robust in so far as they were widely replicated across laboratories.

[19] This of course affords the possibility that the normal rats in the standard hidden condition could use both strategies to locate the platform (see for example Brandeis et al. 1989; D'Hooge and De Deyn 2001).

To return to the framework I introduced in Sect. 3 of this paper, investigators in cognitive neurobiology occupy different perspectives with respect to what experimental paradigms may be used to circumscribe. Some, like Sweatt, use learning paradigms exclusively to draw conclusions about the role of cellular and molecular activity in the production of observable changes in behavior. Thus, one is not likely to find much in the way of an analysis of "what" and "how" rats trained in the hidden condition learn (see Sullivan forthcoming). Furthermore, how a set of observable changes in behavior is classified will either not matter so much, or it will only matter in so far as one is attempting to correlate stages in synaptic, cellular and molecular changes with behavioral ones. Other cognitive neurobiologists, like Dudai and Morris, who acknowledge that there are representational changes that mediate behavioral changes are more likely to point out patterns in the behavioral data that indicate the formation of discrete kinds of representations, although they are not expressly interested in the project of experimentally teasing them apart and investigating them. Finally, it turns out that investigators who are interested in the component cognitive processes that contribute to observable changes in behavior discriminated in learning paradigms like the water maze tend to work in more cognitive areas of neuroscience, which may also be appealed to as a factor contributing to the oscillations in cognitive neurobiology as to how to classify what is under study in the water maze.

Given that some cognitive neurobiologists have sought to better understand what is under study in the water maze, various options as to what the functions are, what is learned and how are on offer across the experimental and review literature and I think this may be cited as one of the reasons why most recently "water maze performance" and "spatial navigation" are used more frequently to capture the observable changes in behavior (see for example, Redish 1999, 2001; Sutherland and Hamilton 2004). For example, rodents trained in the maze are said to learn: to swim (Morris 1981), to swim towards the platform, to swim away from the side walls, that escape is possible or that the platform offers escape (Morris 1981), how to climb on top of the platform or use the platform (Morris 1989, 3046), where the platform is in relationship to distal room cues (Morris 1981), where the platform is in relationship to distal cues and where they are ("allocentric representation") (Eichenbaum et al. 1990), associative relationships (e.g., swim (R) towards that cue configuration (S)), and how to navigate the maze (e.g., Brandeis et al. 1989; D'Hooge and De Deyn 2001; Morris and Frey 1997). The forms of learning include: place learning, associative learning, reward-based or reinforcement learning, spatial learning and non-spatial learning (see for example Hodges 1996). The learning strategies are thought to include: snapshot strategies, taxis strategies, mapping strategies and praxis strategies (e.g., Redish 1999). The navigational strategies are said to include: place navigation, taxon navigation, typically classified under the more general category of spatial navigation (e.g., Redish 1999, 2001). Other functions involved include vision, sensory perception, attention, working memory, long-term memory (i.e., encoding and retrieval), and the activity of the entire motor system.

Although I have not included in the list the processes involved during probe trials for spatial bias, I think two points have been made. The first is that there are a lot of observable changes in a rodent's behavior in the water maze to which an investigator may appeal in order to understand those cognitive processes internal to the organism

that are involved in different stages of the task. The intention of providing the list is to make the case that the behavioral effects treated independently or together correspond to or are by-products of various kinds of representational changes, cognitive capacities and processes. When we think about 'what' the maze circumscribes from this perspective, classifying the behavioral effects under "spatial learning" or "spatial memory" serves to wrongly lump together the many processes that likely contribute to these behavioral effects. To classify them under "water maze performance" (e.g., D'Hooge and De Deyn) avoids this problem, but this solution will not be satisfying for an investigator who is not aiming to discover the component cognitive functions or cellular and molecular mechanisms of water maze performance. I take the vast majority of cognitive neurobiologists as intending to direct their mechanistic explanations at something more general. I think this explains the recent tendency to classify the behaviors under study in the water maze and other maze tasks more generally as "spatial navigation", which would be inclusive of all of the aforementioned kinds of changes in internal representation and componential processes and serves to link those changes and processes to observable changes in behavior in a way that "spatial learning" and "spatial memory" do not. Of course, given the lack of control over how and what is going on with respect to animals trained in the hidden condition of the water maze, perhaps, as D'Hooge and De Deyn indicate, we should either develop better experimental paradigms or as Sutherland and Hamilton indicate, we should turn our attention towards providing a more systematic account of the component informational processes that come online in such training contexts.

6 Conclusions

My first aim in this paper was to address the question of what phenomenon the Morris water maze circumscribes. The generating circumstance was that the philosophical, experimental and review literature provided no guidance in locating an answer to the question. The second aim of the paper was to identify some of the potential sources of the oscillations in those concepts used to refer to the changes in behavior delineated by the water maze across the history of its use. I do not take myself to have offered anywhere close to a thorough analysis, although I think the preliminary analysis does successfully identify some of the sources.

I think the kinds of fluctuations that occurred in the experimental and review literature with respect to what the water maze delineated were to some extent unavoidable in the 1980s and 1990s given the drive in cognitive neurobiology towards identifying the cellular and molecular mechanisms of learning and memory. However, as neuroscientists Robert Sutherland and Derek Hamilton (2004) suggest, investigators who use experimental devices like the water maze can no longer ignore "the constituent [informational] processes" (688, 697) that "contribute to their behavioral outcome measures" (688). They posit a model intended to serve as a framework for thinking about the nature of these processes, thus stressing the importance of functional analysis (see for example, Cummins 1975) for getting clear on what phenomena experimental learning paradigms actually circumscribe. In a similar vein, I suggest (Sullivan forthcoming) that if investigators in cognitive neurobiology are interested in providing

cellular and molecular explanations of more than behavioral performance in this or that experimental learning paradigm, identifying how and what organisms trained in them actually learn will have to be, going forward, a fundamental part of the research project. The precise details of how this will work in practice will have to be left for another paper.

Acknowledgments An earlier version of this paper was presented at two philosophical workshops. The first workshop entitled "Realization, Multiplicity and Experimentation in Biology, Psychology & Neuroscience" was organized by the author and held at the University of Alabama at Birmingham in February 2010. The author benefitted greatly from comments by audience members including: Marshall Abrams, Ken Aizawa, Erik Angner, Ted Benditt, John Bickle, Carl Gillett, Philippe Huneman, Tom Polger, Larry Shapiro, Lynn Stephens and Sven Walter. The author would also like to thank Harold Kincaid and *The Center for Ethics and Values in the Sciences* for generously supporting this workshop. The second workshop was "Current Topics in Philosophy of the Human Sciences" organized by Uljana Feest and held in June 2010 at the Technische Universität in Berlin, Germany. The author would like to thank Uljana for the opportunity to present this paper and members of that audience, especially Maria Kronfeldner and Catherine Stinson, for helpful comments during the discussion period. Finally, the author would like to thank Floh Thiels for many very helpful discussions about on the water maze.

References

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Lawrence Erlbaum.

Bechtel, W. (2009). Molecules, systems, and behavior: Another view of memory consolidation. In J. Bickle (Ed.), *Oxford handbook of philosophy and neuroscience* (pp. 13–40). New York: Oxford University Press.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht: Kluwer Academic Publishing.

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese, 151*, 411–434.

Bliss, T., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology, 232*(2), 331–356.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*, 303–352.

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*, 29–69.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. New York: Oxford University Press.

Chang, H. (2009). Operationalism. *Stanford Encyclopedia of Philosophy Online*. Retrieved July 16, 2009, from http://plato.stanford.edu/entries/operationalism/.

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C., & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P. K. Machamer, R. Grush, & P. McLaughlin (Eds.), *Theory and method in the neurosciences* (pp. 112–136). Pittsburgh: University of Pittsburgh Press.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*(20), 741–765.

Davis, S., Butcher, S., & Morris, R. G. (1992). The NMDA receptor antagonist D-2-amino-5-phosphonopentanoate (D-AP5) impairs spatial learning and LTP in vivo at intracerebral concentrations comparable to those that block LTP in vitro. *Journal of Neuroscience, 12*(1), 21–34.

D'Hooge, R., & De Deyn, P. (2001). Applications of the Morris water maze in the study of learning and memory. *Brain Research Reviews, 36*, 60–90.

Dudai, Y. (1989). *The neurobiology of memory: Concepts, findings, trends*. Oxford: Oxford University Press.

Dudai, Y. (2002). *Memory from A to Z: Keywords, concepts and beyond*. Oxford: Oxford University Press.

Eichenbaum, H., Stewart, C., & Morris, R.G. (1990). Hippocampal representation in place learning. *The Journal of Neuroscience, 10*(11), 3531–3542.

Feest, U. (2003). *Operationism, experimentation, and concept formation*. Doctoral Dissertation, University of Pittsburgh.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences, 41*(2), 131–149.

Feest, U. (Forthcoming). What exactly is stabilized when phenomena are stabilized? *Synthese*.

Hebb, D. O. (1949 [2002]). *The organization of behavior*. Mahwah, NJ: Lawrence Erlbaum.

Hodges, H. (1996). Maze procedures: The radial-arm and water maze compared. *Cognitive Brain Research, 3*, 167–181.

Machamer, P. K. (2009). Neuroscience, learning and the return to behaviorism. In J. Bickle (Ed.), *The Oxford handbook of philosophy and neuroscience* (pp. 166–178). Oxford: Oxford University Press.

Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*, 239–260.

Morris, R. G. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods, 11*, 47–60.

Morris, R. G. (1989). Synaptic plasticity and learning: Selective impairment of learning rats and blockade of long-term potentiation in vivo by the N-methyl-D-aspartate receptor antagonist AP5. *Journal of Neuroscience, 9*(9), 3040–3057.

Morris, R. G. (1990). Toward a representational hypothesis of the role of hippocampal synaptic plasticity in spatial and other forms of learning. *Cold Spring Harbor Symposium in Quantitative Biology, 5*, 161–173.

Morris, R. G. (2003). Long-term potentiation and memory. *Philosophical Transactions of the Royal Society of London B, 358*, 643–647.

Morris, R. G., Anderson, E., Lynch, G., & Baudry, M. (1986). Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature, 319*, 774–776.

Morris, R. G., Davis, S., & Butcher, S. P. (1990). Hippocampal synaptic plasticity and NMDA receptors: A role in information storage? *Philosophical Transactions of the Royal Society of London B Biological Sciences, 329*(1253), 187–204.

Morris, R. G., & Frey, U. (1997). Hippocampal synaptic plasticity: Role in spatial learning or the automatic recording of attended experience? *Philosophical Transactions of the Royal Society of London B, 352*, 1489–1503.

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

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence form unit activity in the freely-moving rat. *Brain Research, 34*, 171–175.

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

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

Redish, A. D. (2001). The hippocampal debate: Are we asking the right questions?. *Behavioral Brain Research, 127*, 81–98.

Restle, F. (1957). Discrimination of cues in mazes: A resolution of the 'place-vs.-response' question. *The Psychological Review, 64*(4), 217–228.

Sullivan, J. A. (2007). *Reliability and validity of experiment in the neurobiology of learning and memory*. Dissertation, University of Pittsburgh.

Sullivan, J. A. (2009). The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuro science. *Synthese, 167*(3), 511–539.

Sullivan, J. A. (Forthcoming). *A role for representation in cognitive neurobiology*. Philosophy of Science, PSA 2008 symposia papers.

Sullivan, J. A., Machamer, P. K., & Thiels, E. (2004). *The study of learning and memory then and now: evidence for conceptual change?* Paper presented at the Society for Neuroscience Annual Meeting, San Diego, CA.

Sullivan, J. A., Machamer, P. K., & Thiels, E. (2005). *The study of learning and memory then and now: Conceptual problems and experimental limitations*. Paper presented at the Learning and Memory Workshop, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sutherland, R. J., & Hamilton, D. A. (2004). Rodent spatial navigation: At the crossroads of cognition and movement. *Neuroscience and Biobehavioral Reviews, 28*, 687–697.

Sweatt, J. D. (2009). *Mechanisms of memory* (2nd ed.). London: Academic Press.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review, 55*, 189–208.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. II. Place learning versus response learning. *Journal of Experimental Psychology, 35*, 221–229.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1947). Studies in spatial learning. V. Response learning vs. place learning by the non-correction method. *Journal of Experimental Psychology, 37*, 285–292.