Concluding Remarks
SFB 1315 Memory consolidation

The Value of a Comparative Approach

Learning from experience is a property embedded into the survival strategies of all animals living in natural surroundings. Animal species live in different ecological niches, are equipped with different sensory and motor capacities, and communicate differently with other individuals of the same species and with other species. They also come with different nervous systems, which can be large or small, and some are highly centralized, while others have several rather separate ganglia. There is no one model animal for this research endeavor, and different animal species have different advantages in searching for the underlying neural mechanisms. One species can be reared more easily in the lab, another has a better-worked-out toolbox of molecular genetics, yet another provides for large and identifiable neurons or allows recordings from multiple neurons simultaneously over long periods of time, while another has already been analyzed in a wide range of behavioral tests. All these, and other practical circumstances, account for and justify the selection of an animal species for study, but the limitation applying to every single species would be disastrous if used to judge the goal of learning and memory research. The main goal of this research is to unravel the general rules and species-specific adaptations in selecting relevant information, adding it to existing knowledge, storing it such that passing time does not eliminate it, and making it available for better-adapted behavioral acts in the future. Comparative studies provide the tool for identifying generalities and specificities. Observing animals in their natural habitat can suggest relevant research questions. While we have ideas about what is worth observing and measuring, we need to be open-minded about unexpected outcomes, as these are often the discoveries that propel research. Food-storing behavior in birds and mammals, communication via gestures and/or sound or by ritualized movements in bees, learning during courtship in *Drosophila*, and navigation are examples. Each of these species provides us with the opportunity to discover novel ways of solving similar problems brought about by an ever-changing environment and to unravel general strategies by comparison.

The transition from the natural habitat to the laboratory is an essential step in hypothesis-driven behavioral research, but is by no means a simple step. We cannot expect to get full control over the animal, which is sometimes a misleading assumption in some of the behavioral studies. However, the history of an individual's experience can be traced more accurately in the laboratory, and the proper control experiments can be established.

Animals need to be constrained for physiological measurements, or they are genetically manipulated to isolate cellular and neural network components of neural function. It is essential to remember that constrained animals or genetically manipulated animals are no longer the living creatures that we observed in their natural environment, and they are not even those seen in the laboratory behavioral tests. It is true that in most cases we do not yet have alternatives attempting to relate neural function with behavioral performance, but while presently there are no better experimental tools of neural recording, we must not forget the distance between the natural conditions and the experimental surroundings where we collect data.

Invertebrates are of particular value in laboratory settings since the transitions from natural to laboratory to experimentally interfering conditions presumably impact them less. Intermediate transition steps from natural to laboratory conditions can be made more easily, and the behavior of these invertebrates appears to be controlled more strongly by innate components.

Nevertheless, transgenic nematodes and flies are not normal animals with just one isolated function that has been modified. It is, therefore, very advantageous that transgenes in *Drosophila* can be switched on and off rather quickly, and even more important, can be genetically rescued, which allows us to test the isolation of the targeted effect very carefully.

Many questions about learning and memory cannot be moved into the laboratory, and these may often be the particularly interesting questions. This has two consequences: (1) The data are correlational in nature as control groups often cannot be studied or serve as partial controls and animal manipulations are very difficult or impossible, and (2) recording brain functions is difficult or impossible. These limitations should not reduce our efforts to collect data under natural conditions, as these data are essential for future laboratory studies and for comparative studies in humans.

A comparative approach should include human beings, and the motivation of many animal studies is to better understand humans. This is justified if appropriate caution is taken and the general limitations of a comparative approach are observed. Both ethology and behaviorism

carry their historical burdens regarding inappropriate generalizations between animals and humans, but cognitive neuroscience offers tools and strategies that help to guide such comparisons. If processes and mechanisms have been identified that apply across animal species, they are less likely to be species-specific adaptations and can safely be generalized to humans. The involvement of phylogenetically homolog brain structures for related forms of learning and/or memory formation are strong hints for homolog functions. The hippocampus (in the case of spatial learning and episodic-like memory) and the amygdala (in fear learning) are two examples. Comparison between animals with very different brain structures (e.g., mammals and insects) is much more difficult, and often no more than analog functions can be assumed. One of the most important and controversial issues related to comparison between animals and humans relates to language and self-awareness. The neural requirements of self-awareness exist in animals but it is not clear whether additional neural functions are required for the human form of self-awareness. The case of episodic memory is a particularly interesting example because essential features of knowledge about what happened, when it happened, and where it happened exist at different degrees of complexity in many animal species. Food-storing birds appear to relate these memories to themselves and appear to expect to find food at that location in the future showing a capacity that is close to personal recollection in humans. Calling this memory "episodic-like" recognizes a gradual, rather than a principle, difference with the introspective experience of mental time travel in humans. This pragmatic approach might be exemplary in the sense that other human mental functions could also be broken down into additive features, which could then be tested for their existence in animals in various combinations and complexities. However, the demonstration of the existence of the components does not prove that the full function of a cognitive faculty as observed (or personally experienced) in humans exists in a particular animal species. Nevertheless, the strength of this approach lies in the assumption that there are no categorical differences between animals and humans, and gradual differences can be traced to different performances according to the complexity of the elements found. An example could be dance communication in honeybees. The bee communicates a location, and depending on the context, the dance might indicate a feeding place, a water or resin resource, or a new nest site. Although the communication process is symbolic and has a vocabulary (although a very reduced one) and a form of syntax (context-dependence), it does not qualify as a language because it lacks essential features, for example, semantics and grammar. One might call it

language-like, as one might categorize other symbolic indexical forms of communications, but the point is that a research program can be set up by this decomposition strategy which allows scientists to search for the related neural processes of the components rather than the mental faculty as a whole.

Theories, Processes, and Mechanisms

Animal learning theory has been a rich research area over the last 90 years or so, and we may ask whether some of its concepts might join with physiological studies for a better understanding of the underlying processes. Theories derived from associative forms of learning have been elaborated the most, and it appears that three concepts are most useful in a search for functional implementations: associative strengths, associability, and prediction error.

Associative strength between two elements (stimulus and/or response) depends on the history of experience and the stimuli/responses involved and controls both acquisition and retrieval of memory. Although different behavioral theories compete for the best way of capturing the essence of associative strengths. Neuroscientists are more than prepared to absorb this concept and translate it into processes of neural plasticity. Donald Hebb (1949) proposed such a neural implementation, and it is widely accepted that synaptic strength is closely related to associative strength. Long-term potentiation and long-term depression are processes that are based on the accurate timing of neural activity in the pre- and postsynaptic elements of neural nets. The coincidence of spike activity as a means of modulating synaptic efficiency appears to play a role not only between pairs of pre- and postsynaptic neurons, but also in networks of many neurons. Coherence of spike activity is an essential feature of cortical nets in up- and downregulation of learning-related neural plasticity. It will be important to show that spike synchrony in biological networks is an emergent property similar to artificial networks and to establish the causal relationship between these global network characteristics and learning. Since small networks composed of identified neurons do not depend on spike coherence in a global sense to establish associative changes in synaptic efficacy (e.g., in mollusks), it will be interesting to search for additional qualities of synchronizing neurons. Such additional qualities could lie in the fact that the three components of memory (acquisition, consolidation and retrieval) are so tightly connected that only under conditions of synchronized activity are all three memory components

activated. New memory contents can only be stored in distributed brain regions which jointly reorganize the network according to the new information, a concept supported by studies on reconsolidation of already stored memory.

Associability is another concept developed in behavioral learning theory that promises to be useful in neural studies. The concept captures the properties of the stimuli and/or outcomes that determine associative strengths as they are reflected in the salience of the stimulus, the predictability or surprise value of a stimulus, or the outcome. Cognitive dimensions of operant learning or perceptual learning involve attention as a critical parameter of learning, a parameter that can be traced to particular structures (e.g., cholinergic projections from basal ganglia, amygdala, and the septohippocampal system).

Prediction error: Learning theories state that learning occurs as long as the outcome of a behavior is not fully predicted, and thus the deviation of the expected from the experienced outcome changes the current associative strengths. Behavioral theories differ with respect to their assumption of whether the error affects associative strength directly.

The implementation of the prediction error into machine learning (Sutton and Barto, 1990) has been very successful, and strong neural correlates exist: for example, the neural properties of reward neurons (dopamine neurons of the mammalian ventral tegmentum (Schultz, 2006) and octopamine VUMmx1 neurons in the honeybee brain (Hammer, 1993; Menzel and Giurfa, 2001).

Forty years ago, Kandel and Spencer wrote a seminal paper entitled “Cellular neurophysiological approaches in the study of learning,” calling for a novel approach in translating basic psychological concepts of learning into strategies for the search for their neural implementations (Kandel and Spencer, 1968). Less than 20 years later, Hawkins and Kandel (1984) presented a first review on their finding on *Aplysia* associative and nonassociative learning and derived neural components comprising a cellular alphabet of learning. This strategy has turned out to be most successful in localizing in space and time neural events induced by learning. It appears that the associative events are distributed, multifaceted, and dependent both on innate predispositions and earlier learning. *Drosophila* provides a particularly carefully

studied case. Different neural structures are involved in learning the same odor by reward or punishment, and short- and long-term memories of the same content reside in different neural nets. Localizing the memory trace is an important step in a functional analysis, and the recent developments in knocking-out, reactivating, recording and stimulating selected neurons in identified networks involved in acquisition, memory formation and retrieval lead to a great step in current cognitive neuroscience (Aso and Rubin, 2016).

A major unresolved issue in both behavioral and neural studies is the relationship between learning with and without external reinforcing or evaluating stimulus. As pointed out concise behavioral theories have been developed for Pavlovian and instrumental conditioning, but perceptual learning, navigational learning, and interval learning provide cases in which no obvious external reinforcer may be present. Is associative learning a special case of a more general form of learning, or is every kind of learning associative? Does an internal reinforcer provide the evaluating function in the latter forms of learning? Learning theory has not settled the debate, and it might well be that functional analysis will show that internal reinforcing circuits are active at the proper time when animals learn by observation. An important component in such forms of learning is attention, but what is the rewarding nature of attention?

Only selectively attended stimuli are learned. Most importantly, modulatory circuits that appear to be involved in coding evaluating stimuli also participate in selective attention. It will be necessary to build conceptual bridges between the concept of associability as developed in theories of associative learning and the evaluating property of directed attention as described in observational learning. Further advances will only be made with the combination of behavioral and neural approaches.

What Is Memory and What Is a Memory Trace?

The many facets of memory are reflected in the many terms used to capture them. Are there 256 different kinds of memory, as Tulving (1972) asked? Irrespective of whether we divide up memories according to time, cellular mechanisms, brain structures involved, categories of contents, type of learning, or type of retrieval, we always imply that memory directs behavior via the process of retrieving information. Brains are equipped with information before, and independent of, acquired information. Thus the content of memory provides a knowledge base

for behavioral guidance (including perception, planning, expecting, and thinking), and splitting it up may obscure the basic and unifying property of memory. One question that needs to be asked, then, is: How do we go about measuring the knowledge stored in memory? We do not know, and this ignorance might be one of the reasons why so much emphasis is placed on the need to define memory by retrieval processes. As long as measurement of memory content is based only on retrieving it from memory, we will not be able to separate stored memory from used memory.

Since the process of memory formation is not directly accessible to behavioral studies, it has been seriously questioned from a behavioral analytical perspective, in terms of whether it makes sense to distinguish between memory as an entity independent of retrieval. The notion of a physical memory trace, independent of its use, however, is a central presumption in neuroscience. Indeed, only when neurologically related interference procedures were introduced into memory research did a clear separation between memory formation and memory retrieval become possible. The key discovery in this context was the consolidation process.

Does a memory exist if it is not retrieved? If the knowledge stored in memory does not guide behavior, a behavioral biologist cannot know whether memory exists (and may thus define memory by its retrievability). But a neuroscientist cannot help but assume that the knowledge stored in memory continues to exist during time periods when it is not retrieved, because the physiological measures of memory are independent of whether the animal performs the corresponding behavior. The concept of memory consolidation is essential in this debate. Hermann Ebbinghaus (1964) described a fast and a slow component in forgetting, and William James (1890) proposed that these may be related to two forms of sequential memories: Primary and secondary memory. The concept of consolidation as a time-dependent process following learning was introduced by Müller and Pilzecker (1900) on the basis of their finding that new learning interfered with the formation of recently acquired memory for short, but not for long intervals. At this stage of analysis, a separation between an internal, time-dependent, and self-organizing process of memory formation and retrieval of memory was not possible, but when experimental interference was introduced and neurological cases of retrograde amnesia were analyzed, strong arguments in favor of an independent engram-building process could be presented. However, the situation is not as simple as was believed. For example, amnesia-inducing procedures could have led to competing learning processes. Irrespective of the

unresolved questions in separating memory formation and memory retrieval processes, the body of evidence is overwhelming, proving that neural traces are indeed induced by the learning process independent of retrieval, and consolidation has a physical basis in the structuring and restructuring processes of neural net properties.

Procedures interfering with ordered neural activity or cellular metabolism during periods of consolidation induce retrograde amnesia. Memory gets better over time, even when it is not used. Sleep phases strengthen the consolidation process, and are related to repetition of content-specific patterns of neural activity.

It appears to me that the debate about the nature of the memory trace will continue as long as we cannot read the encoding processes and directly measure knowledge stored in neural nets. Once we can show these in suitable animals such as *Drosophila*, we will probably discover that, in addition to the constructive processes of reactivating memory and using its content, there is an essential component that exists independent of the reactivation process. Whether we like to call this lasting component memory is a question of definition.

Reactivation of memory leads to new learning and its subsequent consolidation processes. Only recently has neuroscience become interested in the mechanistic aspects of extinction learning and memory formation. The phenomena subsumed under the term reconsolidation provide case studies. Reconsolidation refers to the effect that retrieving memory may lead to cue-dependent amnesia if the retrieval process is followed by treatment with an amnestic agent. What are these learning and reconsolidation processes? Does reactivation indeed make the old memory trace vulnerable to amnestic interference, indicating that new learning overwrites old memory, or do the learning processes involved in memory reactivation induce parallel consolidation processes that reflect the addition of a new memory trace to the existing one? The ongoing debate reflects the same dilemma addressed above. Our inability to measure knowledge as stored information directly restricts our mechanistic analysis to global and indirect arguments. Once again, behavioral analysis needs to be combined with fine-grained neural analysis addressing the critical question much more directly at the level of the neural elements of the engram.

What might be a suitable strategy toward a direct reading of knowledge? A first step has been already highly successful in particular in Drosophila and mice, identifying and localizing neurons that are essential for the acquisition processes. Neurons or a subset of a neural net were also determined that are required for retrieving a particular memory. Furthermore, in a few cases

neurons were identified that are necessary and sufficient to shift early forms of memory into stable forms, and neural synchrony between them seems to play an essential role.
However, the procedures do not yet provide us with access to information stored in the memory trace but rather capture (only) the processes involved in building and/or retrieving it. Possibly one needs the whole nervous system of the animal in question to recall the neural conditions that have led to all the changes necessary to store the content of the memory. A large and powerful battery of highly sophisticated molecular–genetic tools are available to measure the spatial–temporal patterns of dynamic changes in selected neurons and neural nets of the *Drosophila* and mouse brain. Reading the dynamics of the neural elements during the learning process (i.e., consolidation and retrieval under conditions in which the animal tells us via its behavior whether it perceives, attends, and retrieves) will guide us but understanding at least part of the knowledge stored in memory requires knowledge reading, a capacity still in the future.

1.01.5. The Engineer's Approach to Learning and Memory

Engineers compose and biologists de-compose, so a combination of these two strategies should be favorable to the study of a complex system such as the brain. Constructive thinking in theoretical neuroscientists is inspired by rules derived from behavioral studies (e.g., Hebb's rule), by the morphology of brains and the connectivity patterns of neurons (e.g., the matrix-like connectivity in the hippocampus), by the functional properties of neurons (e.g., synaptic plasticity), and by theoretical concepts developed independently from, but motivated by, thoughts about how the brain might work (e.g., autoassociative or attractor networks).
Irrespective of the intellectual pleasure one experiences when thinking about theoretical neural nets, one might ask how the joint efforts propel our understanding. I see the following three points:

1. Hypothesis-driven research like ours requires well-formulated concepts and hypotheses. Theories developed for neural nets shape these concepts and allow us to formulate predictions.

2. The analysis of the vast amounts of data collected by anatomical, electrophysiological, opto.-physiological, and molecular studies requires the contribution of theoretical neuroscientists to extract relevant information and interpret it.

3. There exists no concise theory of the brain. Global brain functions need to be constructed from elemental and network functions and implemented into a model.

At any of these levels of a modeling approach, one has to decide what is considered an essential feature and which of the many characteristics of the neurons, their connectivity at the local and the global level, are implemented or not. Should one use simplified integrate-and-fire neurons or Hodgkin-Huxley-type neurons? Should the model care about the real gestalt of neurons or not? How seriously should one take the neuroanatomical data on local and global connectivity in particular since the full connectivity of a whole brain (*Drosophila*) is now at reach? These and many other decisions are hard to make, and different choices produce serious debates about the suitability of these models. There are many measures of suitability: Are experiments stimulated, predictions offered, and interpretations of data supported or rejected?
Ultimately, models of neural function should also predict behavioral outcomes. It is to be expected that the success of the combined theoretical and experimental approach will make modeling an indispensable part of the search for the memory trace.

Conclusion

Curiosity-driven behavioral studies, theory-guided laboratory behavioral experiments, and modeling of neural functions define a unique workspace in the search for the engram. Joining forces will help, and the research projects of this SFB will (hopefully) facilitate communication between these disciplines. The task is indeed demanding, because the goal will not only be to localize and characterize the memory trace, but to measure the knowledge stored in the memory trace independent of and in addition to the behavioral read-out process.