

Neurodynamics in auditory cortex during category learning

FRANK W. OHL¹, HENNING SCHEICH¹ & WALTER J. FREEMAN²

¹ Leibniz Institute for Neurobiology, Magdeburg, Germany

² University of California, Berkeley, USA

Ch. 8 in: König R, Heil P, Budinger E, Scheich H (eds.) *The Auditory Cortex — A Synthesis of Human and Animal Research*. Mahwah NJ: Lawrence Erlbaum Assoc. pp. 429-444.

Introduction

Research on learning has for historical reasons been divided mainly into behaviourally oriented studies performed on animals and more cognitively oriented studies in humans (Anderson, 2000). Accordingly, studies aimed at the presumed neurophysiological basis of learning have either exploited the full range of cognitive-phenomenological approaches in humans while being methodologically confined mainly to non-invasive imaging techniques, or alternatively, have made use of the better accessibility of the animal nervous system physiology while being restricted in the definition and analysis of the cognitive aspects involved. Consequently, our physiological understanding of learning processes is best for simple behaviours that can be studied in suited animal models (e.g. Kandel, 2001) but declines for cognitively more demanding aspects of learning.

In the present chapter we present an animal model, amenable to detailed physiological analysis, of a cognitively demanding learning task, namely the formation of categories (concept formation) and the sorting of novel stimuli into these categories. Both of these aspects are encapsulated in the term 'category learning'.

We will first argue that learning phenomena having aspects beyond mere stimulus-response associations, like category learning does, are of high relevance for studying the neuronal basis of learning in general, in that they preclude explanation of learning phenomena by a broad class of simple neurophysiological models which are otherwise discussed as elemental for physiological theories of learning. Second, we introduce a new animal model for studying category learning and describe an electrophysiological correlate of category-specific processing of stimuli. Third, we argue that the physiological results derived from the category learning paradigm contribute to solving an important problem reported in the literature since the 1950s. This is the problem that in trained animals activity patterns in sensory cortices, though often stable and identifiable with high significance, do not seem to be invariant with stimuli or even with a particular training context, as required for a 'representation of a stimulus' by classic sensory physiology (see Freeman, 2000, and references therein).

What is category learning ?

We define category learning as the process by which categorization of stimuli, or, more generally, of situations experienced by a subject is acquired. The term categorization describes the phenomenon that under suited conditions a (human or non-human) subject might behaviorally respond to a multitude of stimuli or situations with a considerably smaller number of response types although the stimuli itself could in principle be discriminated by the subject. (For a recent discussion of the relationship between category learning and category use see Ashby & Ell, 2001, and Markman & Ross, 2003). The nature of categorization has

been subject to intellectual debates at least since Aristotle. A brief review of some of the more traditional problems which have repeatedly emerged in the context of characterizing the nature of categorization will be used to motivate the viewpoint that has been taken in planning the experiments focused on in the present chapter.

Traditional accounts of categorization

The Aristotelian view (often referred to as the 'Classical Theory') considers having or establishing a set of necessary and/or sufficient criteria to be met by stimuli or situations as being the essence of determining their membership to a category. This view, enriched by an appropriate formal framework, is also maintained by some 'artificially intelligent' approaches to the categorization problem. A main objection that was brought up against this view is that for many 'natural' categories such sets cannot in fact be found (e.g. Rosch, 1973): Neither seems category membership always be defined by features which can unequivocally be attributed to all members, nor can features always be evaluated to indicate membership or no membership to a category. Rather, varying degrees of 'typicality' of features and/or category members seem to be the rule. Consequently, a number of accounts have been proposed which can be summarized under the heading of 'prototype theories' (e.g. Posner & Keele, 1968). Prototype theories hold that some form of idealized representation of category members (a Platonic ideal form) exists and actual membership of a given stimulus is determined by some scaling of 'similarity' of this member to the prototype. Prototype theories in some instances have faced the problem of providing a convincing framework for establishing the required similarity relations (for discussion see Ashby & Perrin, 1988) or could not account for the phenomenon of non-prototype members having more pronounced effects on categorization performance than the prototype itself (e.g. Brooks, 1978; Medin & Schaffer, 1978). Both studies on human (Rosch, 1975) and animal categorization (Lea & Ryan, 1990) have argued that some of these problems can be attenuated by lifting the requirement for local comparisons with a singular prototype and instead requiring global comparisons with multiple (in the extreme form with all) category members (Estes, 1986; Hintzman, 1986; Nosofsky, 1986) for which reason such theories are summarized under the heading 'exemplar theories'. While historically these (and other) approaches have been initiated and put forward by their authors with very different rationales in mind they can be more objectively compared to each other using a suitably formalized reformulation (Ashby & Maddox, 1993) which can be derived from general recognition theory (Ashby & Townsend, 1986). It is particularly disturbing that many experimentally achievable observations can be accounted for by more or all of the theoretical viewpoints provided, at least if appropriate modifications to the extreme positions of such theories are allowed (Pearce, 1994).

Categorization in non-human species

While in human studies the above sketched (and other) viewpoints are maintained without questioning the existence of concepts, as mental constructs, as a possible basis for categorization, this option is not so clear for non-human species, simply because they cannot report on having a concept that might guide their categorization behavior. Research on categorization in animals has therefore to rely on operational criteria. It has been argued that

defining a concept would require reference to language so that nonhuman species could not have concepts at all (Chater & Heyes, 1994). A less extreme viewpoint suggests that some instances of animal categorization might reflect mere discriminations albeit with very complex stimuli (Wasserman & Astley, 1994). It has been elaborated that the demonstration of concepts in non-human species might be possible by suitably designed transfer experiments with appropriately constructed control stimulus sets (Lea, 1984). In designing category learning experiments for non-human species care must be taken because natural and even artificial stimuli could already have ecological relevance for the species given causing a bias in its learning behavior (Nelson & Marler, 1990). In any case, the transfer of learned behaviors to novel stimuli is a relevant criterion because it demonstrates that during an instantiation of category learning more has happened than mere associations of responses to trained particular stimuli (Lea & Ryan, 1990).

Categorization and generalization

It is appropriate to consider the terms 'categorization' and 'generalization' at this point in more detail as they are inconsistently used in the literature and have a bearing on the experiments to be described in this chapter.

When a (human or non-human) subject is trained to discriminate a stimulus A from a stimulus B the discrimination performance typically develops gradually over some time as is manifest in the various forms of functions usually referred to as 'learning curves'. A typical form is schematized in Fig. 1 A which displays the temporal evolution of the hit rate and false alarm rate in a GO/(NO-GO) discrimination experiment. Other depictions of the discrimination learning behavior are possible depending on the kind of experiment performed (be these symmetric choice experiments, signal detection approaches, etc.) and the choice of behavioral observables (e.g. the various transformations of hit rate and false alarm rates which are suitable under given conditions). Generally, however, the changing conditional rate of occurrence of some behavior must be assessed at some point in the analysis. The typically asymptotic response rate depends on various parameters (vigilance, internal response biases, strength of the learned association, etc.) but also shows some degree of stimulus specificity. This can be demonstrated by measuring what is called generalization gradients: In a GO/(NO-GO) experiment in which a subject is trained to show the GO reaction in response to stimulus A and the NO-GO reaction in response to stimulus B, for example, a generalization gradient can be assessed by measuring the conditioned response (typically quantified by its frequency of occurrence across trials or by the strength of its expression) as a function of a physical distance between parameters characterizing stimuli A and B. The latter defines a path through the parameter space connecting stimuli A and B. When physical stimulus parameters are varied along this path a more or less gradual fall-off in the A-specific response amplitude is typically observed and referred to as a generalization gradient (Fig. 1 B).

Conversely, when a subject has formed categories it can recognize even novel stimuli as representants of the learned categories. A depiction of behavioural variables analogous to a learning curve would therefore indicate a high discrimination performance even in the first training session (Fig. 1 C). The experimental demonstration of this behaviour is critical for assessing category learning and is sometimes referred to as the criterion of the 'transfer of

learned behaviours to novel stimuli' (see subsection on behavioural results in the next section). Therefore, category learning is distinguished from simple or discriminative conditioning also by the psychometric functions it produces. Instead of gradual generalization gradients we find sigmoid psychometric functions with a more or less sharp boundary at some location in the stimulus parameter space, called the categorization boundary (cf. Ehret, 1987) (Fig. 1 D). Learned categories develop as cognitive constructs; they epitomize subjective 'hypotheses' that are expressible as parcellations of the set of actually perceived or imaginable stimuli, conditions or actions, into equivalence classes of meaning in particular contexts. Transfer of learned behaviours to novel stimuli therefore follows the subjective laws of this parcellation, rather than being guided by (physical) similarity relations between stimuli or stimulus features. These are important criteria for cognitive structures that have to be accounted for by physiological models of learning.

In this context it could be noted that in some artificial systems (like artificial neural networks) generalization gradients emerge as a result of the used training and learning paradigm. Already in simple one-layer perceptrons generalization gradients can mimic categorization behaviour when threshold operations are applied to suited state variables of the network. Typically, however, apart from the thresholding operations such systems undergo only smooth transformations under the learning paradigm unlike what has been described for category learning and we prefer to call this behaviour classification. This does not imply however that artificial systems cannot in principle show category learning. It is conceivable that artificial systems can be equipped with the requisite process capacities so that their emergent behaviour could in all fairness be called category learning (Nakamura, et al., 1993; Kozma and Freeman, 2003).

In a nutshell, generalization is a general feature of learned stimulus-cued behaviours reflecting the converse of stimulus specificity, while categorization is a cognitive process based on the parcellation of the represented world into equivalence classes of meaning, valid for an individual in a particular context and in a particular time.

On the relationship between category learning and neuronal plasticity

The exact nature of the relationship between category learning, or learning in general, and neuronal plasticity is highly nontrivial, because the former is a purely psychological concept while the latter is a physiological concept. Conceptual difficulties in inter-relating these two domains are therefore predictably similar to other situations in science where conceptually different levels have to be linked, like in the case of relationship between thermodynamics and statistical physics, or in the case of the 'mind-body problem'. In this section it will be argued that traditional physiological accounts for learning phenomena are insufficient to characterize the physiological basis of category learning.

If the role of neuronal plasticity for learning is considered by physiologists, it is usually seen in the capacity for re-routing the flow of excitation through neuronal networks. The case of Pavlovian conditioning, a learning phenomenon that has been intensively studied by physiologists, lends itself to this concept in a very straightforward way: Pavlovian conditioning can be described, and - in fact - has traditionally been defined, as a process by which an initially behaviourally neutral stimulus can later elicit a particular behaviour, when it has previously been paired with a stimulus (US) that unconditionally triggers this behaviour (Fig. 2 A). It has proved successful to formulate models of the role of neuronal plasticity for learning which basically consist in a one-to-one translation of this idea into a neuronal

substrate (Fig. 2 B). In the case of the conditioned gill withdrawal reflex in *Aplysia*, for example, this concept is manifest in the feedforward convergence of – in the simplest case – two sensory neurons on a shared interneuron which in turn projects on a motor neuron. The concept is so straightforward, that its appraisal as a generic element of physiological models of learning has been put forward, as, for example, expressed by the metaphor of a 'cellular alphabet' (Hawkins & Kandel, 1984) or the metaphor of a 'molecular alphabet' (Kandel et al., 1995).

The claim of the elemental nature of the above sketched neuronal model for learning has been challenged by cognitive science (e.g. Schouten & De Long, 1999) where learning is viewed as a process by which an animal gains information about conditions in its environment that help it to behave in meaningful ways. It should be noted that this perspective includes, among other things, the possibility of meaningful behaviours to novel stimuli. Novel stimuli, however, have not been encountered before, specifically, they have not been associated with unconditional triggers (Fig. 2 C). Therefore, a simple feedforward convergence scheme as in Fig. 2 A cannot be used as an explanation for such aspects of learning, i.e. aspects that go beyond mere stimulus-response associations.

The current chapter focuses on an example of category learning because category learning particularly emphasizes this aspect of learning, i.e. the meaningful behaviour in response to novel and unfamiliar stimuli, and therefore precludes explanation by a broad class of reductionist schemes. In this sense, the study of category learning is of general relevance to the physiological understanding of learning phenomena in general.

Moreover, the formation of categories is fundamental to cognition (Estes, 1994). Category learning transcends simple stimulus-response associations, typically studied in systems neurophysiology, in that it involves abstraction from learned particular stimulus features. Category learning leads beyond the information given (Kommatsu, 1992).

A new animal model of category learning

In this chapter the Mongolian gerbil (*Meriones unguiculatus*) is introduced as a new model of category learning (Wetzel, et al., 1998; Ohl, et al., 2001). The following two sections will discuss its behavioral analysis and present the physiological data obtained during the category learning behavior, respectively. As motivated in the previous subsections, the experiment is designed to shed light on the processes that give rise to the transfer of learned behaviors to novel stimuli which are located outside the generalization gradient in the perceptual space. The occurrence of the transfer of the learned behaviors to novel stimuli will be used as a marker event for the study of the physiological correlates of category learning. With respect to the issues addressed in the subsection on categorization in non-human species a stimulus set was used which is demonstratively behaviorally neutral to the naïve gerbil. Moreover, stimuli were so selected that they did not differ in the ease with which they could be associated with the particular behaviors used in the training experiments. With respect to the background summarized in the subsection on categorization and generalization, care was taken to separate generalization behavior from categorization behavior.

Categorization of modulation direction in frequency modulated tones - behavioral analysis

Stimuli

As stimuli we used linearly frequency-modulated tones traversing a frequency range of 1 octave in 250 ms in rising or falling fashion, played at an intensity of 70 dB SPL as measured at a distance of 10 cm in front of the speaker. Due to reflections of the sound wave in the shuttle box and various possible head positions of the animal during the experiment a considerable variance of effective stimulus intensity across trials is predictable. In frequency modulated tones a number of parameters can be (co-)varied, like their duration, their intensity, the frequency ranges covered, and the modulation rate, i.e. the rate of change of the tones' instantaneous frequency. By construction, most of those stimuli (all, except for those with zero modulation rate, i.e. pure tones) can be categorized as either 'rising' or 'falling frequency modulated tones', depending on whether the instantaneous frequency changes from low to high or from high to low, respectively.

For the experiment stimuli were so designed that they fell outside the generalization gradients established by training naïve animals to the two neighboring (in the parameter space) stimuli. This ensures that observed learning curves would resemble the schemes in Fig. 1 A and C as indicating that the subject has not or has categorized the stimuli, respectively.

Apparatus and training paradigm

Training consisted of a GO/(NO-GO) avoidance paradigm carried out in a 2-compartment shuttle box, with the two compartments separated by a little hurdle, ensuring a low rate of spontaneous hurdle crossings. Animals were trained to cross the hurdle in response to a rising frequency modulated tone and to stay in the current compartment in response to a falling frequency modulated tone. Training was organized in so-called 'training blocks' during which the discrimination of one particular rising frequency modulated tone from a tone traversing the identical frequency range in falling direction was trained. A training block consisted of a number of training sessions, with one session trained per day and was continued until no further changes in conditioned response rates were achieved in three consecutive sessions. Then another training block was initiated in which the discrimination of another pair of a rising and falling frequency modulated tone was trained. A training session encompassed the randomized presentation of 30 rising and 30 falling frequency modulated tones with the animals' false responses (misses and false alarms) being negatively reinforced by a mild electrodermal stimulation through a metal grid forming the cage floor. Control groups were run with the opposite contingencies to test for potential biases in this behavior which have been reported for the dog (McConnell, 1990). We showed that for the stimulus parameters tested no such biases existed for the gerbil.

Behavioral results

Animals trained on one or more training blocks never generalized to pure tones of any frequency (e.g. start or stop frequencies of the modulated tone, or frequencies traversed by the modulation or extrapolated from the modulation). This could be demonstrated by direct transfer experiments (Ohl, et al., 2001, supplementary material) or by measuring generalization gradients for modulation rate which never encompassed zero modulation rates (Ohl, et al., 2001).

Categorization was demonstrated by the transfer of the conditioned response behavior (changing compartment in response to tones of one category and remaining in the current compartment in response to tones of the other) to novel stimuli as measured by the response rates in the first session of a new training block. This sequential design allowed the experimenter to determine the moment in which an individual would change its response behavior from a 'discrimination phase' (Fig. 1 A and C) to a 'categorization phase' (Fig. 1 B and D). All animals tested were able to categorize novel frequency modulated tones but, most notably, different individuals showed the transition from the discrimination phase to the categorization phase at different points in time in their training histories, although all had been trained with the same sequence of training blocks (Ohl et al., 2001). Also, the transition from the discrimination phase to the categorization phase occurred abruptly rather than gradually, i.e. in the first session of a training block either no discrimination performance was observed (discrimination phase) or the full performance was observed (categorization phase). A third property of the transition was that after its occurrence the discrimination performance remained stable for the rest of the training blocks. These three properties of the transition, individuality of the time point of occurrence, abruptness of occurrence and behavioral stability after its occurrence, make it resemble a state transition in dynamic systems. We used this transition as a marker in the individual learning history of a subject to guide our search for physiological correlates of this behavioral state transition.

Physiological correlates of category learning

Neuronal representation of utilized stimuli

A suitable level for studying electrophysiological correlates of perceptual organization is the mesoscopic level of neurodynamics (Freeman, 2000), which defines the spatial scale of phenomena observed (Barrie, Freeman and Lenhart, 1996) and provides focus on electrical phenomena emergent from the mass action of ensembles of some 10^4 to 10^5 neurons. Since it was demonstrated for the gerbil that the discrimination of the direction of frequency modulated tones requires a functional auditory cortex (Ohl, et al., 1999; Kraus, et al., 2002) the training procedures were combined with the parallel measurement of the neurodynamics in auditory cortex. For cortical structures, this level of description is accessible by measurement of the electrocorticogram. We have therefore combined the described category learning paradigm with the measurement of the electrocorticogram using arrays (3 – 6) of microelectrodes chronically implanted on the dura over the primary auditory cortex. The spatial configuration and interelectrode distance (600 μm) of the recording array were so designed to cover the tonotopic representation of the frequency modulated stimuli used and avoid spatial aliasing of electrocorticogram activity (Ohl, et al., 2000a).

The spatial organization of the thalamic input into the auditory cortex can be studied by averaging electrocorticograms across multiple presentations of the same stimulus, yielding the well-known auditory evoked potential (Barth and Di, 1990, 1991). Our studies of pure-tone-induced (Ohl, et al., 2000a) and frequency-modulated-tone-induced (Ohl, et al., 2000b) auditory evoked potentials in primary auditory cortex, field AI, revealed that their early components (P1 and N1) are topographically organized, i.e. are localized at positions within the tonotopic gradient of the field that correspond to the frequency interval traversed by the frequency modulation, while their late components (P2 and N2) are not. On a finer spatial scale, the localization of the early components of rising and falling frequency modulated tones was found to be shifted towards tonotopic representations of the respective end frequencies of the modulations, i.e. towards higher frequencies for rising modulations and towards lower frequencies for falling modulations. These 'tonotopic shifts' (Ohl, et al., 2000b) could be explained by the finding that single neurons are usually activated more strongly when the frequency modulation is towards the neuron's best frequency than when it is away from it (Phillips, et al., 1985). In the former case, the activation of frequency channels in the neighbourhood of the best frequency of a single neuron are recruited more synchronously than in the latter case, due to the increasing response latency with increasing spectral distance from the neuron's best frequency. If this asymmetry is transferred onto a tonotopically organized array of neurons a tonotopic shift as described will result. Tonotopic shifts have previously been reported in the cortex analogue of the chick (Heil, et al., 1992).

Single trial analysis of electrocorticograms

Since physiological correlates of category learning could not be expected to occur time-locked to stimulus presentation we analyzed electrocorticograms recorded during the training with a single trial type of analysis. Instantaneous spatial patterns in the ongoing cortical activity were described by state vectors (Barrie, Freeman and Lenhart, 1996; Ohl, et al., 2001). State vectors were formed from estimates of signal power in 120 ms time windows obtained for each channel. As the spatial pattern of signal power evolved over time the state vector moved through the state space along a corresponding trajectory. For each trial, the Euclidean distance (parameterized by time) to a reference trajectory was calculated and termed 'dissimilarity function'. In each case the reference trajectory was the centroid over trajectories associated with trials associated with stimuli from the respective other category measured in the same training session. I.e., each trajectory associated with a rising frequency modulated tone was compared to the centroid over all trajectories associated falling frequency modulated tones in the same session, and vice versa. Comparison of single trajectories with centroids of trajectories, rather than other single trajectories, ensured that, on a statistical basis, transient increases in the pattern dissimilarity (peaks in the dissimilarity function) were due to pattern changes in the observed trajectory rather than in the centroid. In naïve animals, dissimilarity functions showed a 'baseline behavior' with a sharp peak (2 – 7 standard deviations of baseline amplitude) after stimulus onset. This peak occurred predictably because of the topographically dissimilar patterns (tonotopic shifts) of early evoked responses that rising and falling frequency modulated tones produce (Ohl, et al. 2000b). With learning, additional peaks emerged from the ongoing activity, thus labelling spatial activity patterns in single trials with transiently increased dissimilarity to the reference trajectory indicating a potential relevance for representing category-specific information processing. These patterns were therefore termed 'marked states'.

To test whether marked states do in fact represent processing of category-specific information we analyzed the similarity and dissimilarity relations among them in the entire course of the training. While animals were in their discrimination phases (prior to the formation of categories), we observed that dissimilarities between marked states within categories were of the same order of magnitude than between categories. After an individual animal had entered its categorization phase dissimilarities within a category were significantly smaller than between categories (Ohl, et al., 2001). This indicated the existence of a metric which reflected the parcellation of stimuli into equivalence classes of meaning. This type of metric is therefore different from the known tonotopic, which reflects similarity relations of physical stimulus parameters, namely spectral composition, in that it reflects subjective aspects of stimulus meaning, namely its belongingness to categories formed by previous experience.

The spatial organization of the emerging marked states was analyzed in more detail and compared to that of the early evoked activity (also yielding peaks in the dissimilarity function) by a multivariate discriminant analysis, identifying the regions in the recording area which maximally contributed to the dissimilarity between the observed pattern and the reference pattern (Ohl, et al., 2003b), or identifying the regions which contribute most information about the pattern (Ohl, et al., 2003a).

Conclusions

It was possible to develop an animal model of auditory category learning which demonstrated the formation of categories as a process which three main characteristics in the behavioural data: First, categorization developed abruptly rather than gradually. Second, it developed at a point in time that was specific for each individual subject in its learning history. Third, when categorization had occurred in a subject it remained stable for the rest of the subject's training experiences (unless a change in the reinforcement schedule forces a change in the meaning attributed to stimuli). A process which conforms to these characteristics is sometimes termed 'Aha'-event to indicate a change in the cognitive state of a subject.

The neurophysiological analysis revealed that the process of associating meaning to acoustic stimuli as indicated by an increasing discrimination performance in the behavioural data was paralleled by the emergence of transient activity states in the ongoing cortical activity, that could be identified on the basis of their dissimilarity to patterns found in trials associated with stimuli not belonging to the category. These activity states are the first demonstration of a 'constructive aspect' of neural activity during a categorization event.

It is noteworthy, that spatial patterns of electrocortical activity in single trials were already observed by Lilly in his topographic studies (Lilly and Cherry, 1954). To him it was already apparent that the long lasting dynamics was not just random 'noise', but was better described by 'figures' moving in time and space across the cortical surface. At that time the majority of research programs had already turned to the analysis of averaged data in which such spatiotemporal structures are no longer detectable. The few research programs pursuing analysis of activity patterns in single trials (e.g. DeMott, 1970; Livanov, 1977) had faced a major problem for the interpretation of such patterns: the lack of invariance with the applied stimuli. A large body of data accumulated over the last decades (summarized in Freeman, 2000) showed that such patterns might remain stable when repetitively evoked by sensory stimulation for a certain period in time, but might typically vary with behavioural context, particularly in learning situations when stimuli were associated with particular meanings. This lack of invariance of these patterns with the mere physical parameters of stimuli challenged

their interpretation as 'sensory representations'. The observed metastability of the patterns was hypothesized to reflect context aspects of the stimulation as well as the perceptual history of the individual, and it was inferred that such patterns reflect subjectively relevant cognitive structures (for summary see Freeman, 2000, and references therein, including the application of basin-attractor theory in nonlinear dynamics to the solution of the problem of categorization). The results described above critically confirm this interpretation: The category learning paradigm, first, allows determination of the point in time when a particular cognitive structure (the formation of the categories 'rising' and 'falling') emerges, and second, predicts that the main source of variance in the stimuli (the spectral interval traversed by the frequency modulation) is no longer a relevant feature after a subject's transition to categorization. Consequently, it was found that the dissimilarity between marked states associated with stimuli belonging to the same category was significantly reduced after the transition to categorization, although the physical dissimilarity of the corresponding stimuli was still high, as also reflected in the topographic organization of the stimulus locked peaks in the dissimilarity function and the fact the dissimilarities remained high in individuals that had not yet formed categories.

In this sense, the utilized paradigm and analysis strategy provided an objective (for the experimenter) window of a subjective cognitive structure (that of the animal).

Acknowledgements

This work was supported by the BioFuture grant of the BMBF to F.W.O, and by grants from the Land Sachsen-Anhalt. We thank Brian Burke and Daniela Labra Cardero (Berkeley), as well as Kathrin Ohl and Thomas Wagner (Magdeburg) for technical assistance. The assembly and surgical implantation of electrode arrays and EEG recording were performed in the Department of Molecular & Cell Biology of the University of California at Berkeley under the supervision of the Animal Care and Use Committee and the Campus Veterinarian in accordance with the guidelines of NIH.

Figures with captions

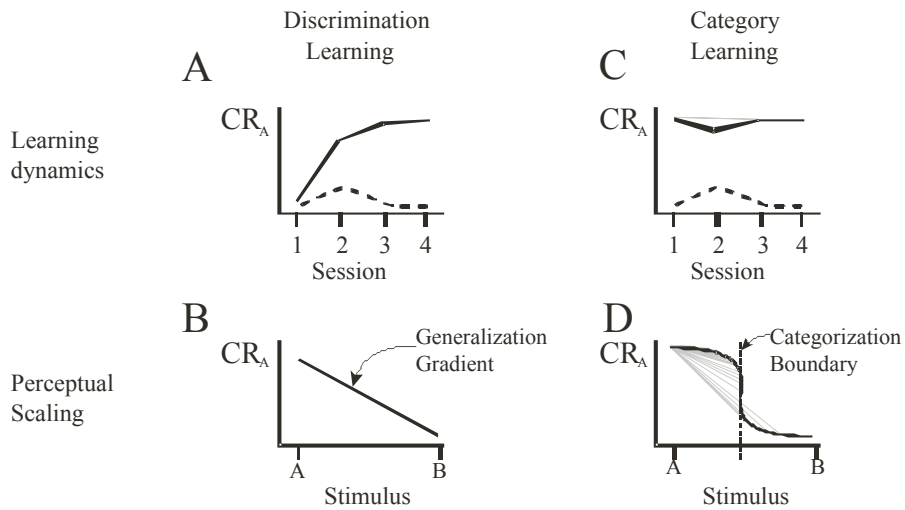


Figure 1. Suitable observables differentiating discrimination learning and category learning. Learning curves (A, C) and psychometric functions (B, D) after discrimination learning (A, B) and after category learning (C, D) are shown. The A-specific conditioned response rate CR_A is typically a measure of the rate of occurrence or of the strength of the response that has been trained to be elicited by stimulus a A. Solid and dashed curves in A and C depict the hit rate and false alarm rate, respectively. See subsection on categorization and generalization for details.

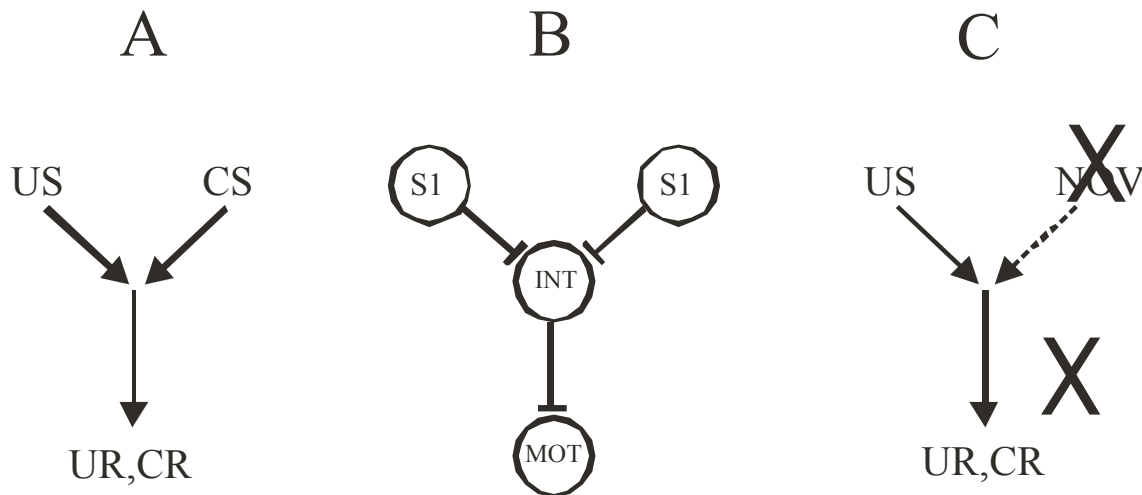


Figure 2. A. General scheme of flow of information before and after Pavlovian conditioning. Before conditioning an unconditioned stimulus (US) will elicit a particular behavior, then referred to as unconditioned response (UR). After conditioning, a previously behaviourally neutral stimulus can elicit this behaviour as a conditioned response (CR) and is then referred to as the conditioned stimulus (CS). B. A straightforward translation of the flow of information during Pavlovian conditioning into a flow of neuronal excitation within a neuronal substrate. C. The architecture in B cannot explain responses to novel stimuli (NOV), because novel stimuli have not been associated with unconditional triggers and, consequently, cannot be conditioned.

References

- Anderson, J.R. (2000). *Learning and memory. An integrated approach.* New York: John Wiley & Sons.
- Asby, F.G., & Ell, S.W. (2001) The neurobiology of human category learning. *Trends in Cognitive Sciences*, 5, 204-210.
- Ashby, F.G., & Maddox, W.T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 7, 372-400.
- Ashby, F.G., & Perrin, N.A. (1988). Toward a unified theory of similarity and recognition. *Psychological Reviews*, 95, 124-150.
- Ashby, F.G., & Townsend, J.T. (1986). Varieties of perceptual independence. *Psychological Reviews*, 93, 154-179.
- Barrie JM, Freeman WJ, Lenhart M (1996) Modulation by discriminative training of spatial patterns of gamma EEG amplitude and phase in neocortex of rabbits. *Journal of Neurophysiology* 76: 520-539.
- Barth, D.S., & Di, S. (1990). Three-dimensional analysis of auditory-evoked potentials in rat neocortex. *Journal of Neurophysiology*, 64, 1527-1636.
- Barth, D.S., & Di, S. (1991) The functional anatomy of middle latency auditory evoked potentials. *Brain Research*, 565, 109-115.
- Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Chater, N., & Heyes, C. (1994). Animal concepts: Content and discontent. *Mind and Language*, 9, 209-247.
- DeMott, D.W. (1970). *Toposcopic studies of learning.* Springfield, IL: Thomas Books.
- Ehret, G. (1987). Categorical perception of sound signals: facts and hypotheses from animal studies. In S. Harnad (Ed.) *Categorical perception* (pp. 301-331). Cambridge, England: Cambridge University Press.
- Estes, W.K. (1994) *Classification and Cognition*, New York: Oxford University Press.
- Estes, W.K. (1986). Array models for category learning. *Cognitive Psychology*, 18, 500-549.
- Freeman, W.J. (2000). *Neurodynamics. An exploration in mesoscopic brain dynamics.* London: Springer-Verlag.
- Freeman, W.J. (2001) *How Brains Make Up Their Minds.* New York: Columbia University Press.

Hawkins, R.D. & Kandel, E.R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Reviews*, 91, 375-391.

Heil, P., Langner, G., & Scheich, H. (1992). Processing of frequency-modulated stimuli in the chick auditory cortex analogue: evidence for topographic representations and possible mechanisms of rate and directional sensitivity. *Journal of Comparative Physiology A*, 171, 583-600.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Reviews*, 93, 411-428.

Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (1995). *Essentials of neural science and behaviour*. Norwalk, CT: Appleton & Lange.

Kandel, E.R. (2001). The molecular biology of memory storage: A dialogue between genes and synapses, *Science*, 294, 1030-1038.

Kommatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, 112, 500-526.

Kozma R, Freeman WJ (2003) Basic principles of the KIV model and its application to the navigation problem. *Journal of Integrative Neuroscience* 2: 125-145.

Kraus, M., Schicknick, H., Wetzel, W., Ohl, F., Staak, S., Tischmeyer, W. (2002). Memory consolidation for the discrimination of frequency-modulated tones in Mongolian gerbils is sensitive to protein-synthesis inhibitors applied to auditory cortex. *Learning & Memory*, 9, 293-303.

Lea, S.E.G. (1984). In what sense do pigeons learn concepts. In H.S. Terrace, T.G. Bever, & H.L. Roitblat (Eds.) *Animal cognition* (pp. 263-276), Hillsdale, NJ: Erlbaum.

Lea, S.E.G. & Ryan, C.M.E. (1990). Unnatural concepts and the theory of concept discrimination in birds. In M.L. Common, R.J. Herrnstein, S.M. Kosslyn, & D.B. Mumford (Eds.) *Quantitative analyses of behaviour*. Vol. VIII. Behavioral approaches to pattern recognition and concept formation (pp. 165-185), Hillsdale, NJ: Erlbaum.

Lilly, J.C. & Cherry, R.B. (1954). Surface movements of click responses from acoustic cerebral cortex of cat: leading and trailing edges of a response figure. *Journal of Neurophysiology*, 17, 531-537.

Livanov, M.N. (1977). *Spatial organization of cerebral processes*. New York: Wiley.

Markman, A.B. & Ross, B.H. (2003). Category use and category learning. *Psychological Bulletin*, 129, 592-613.

McConnell, P.B. (1990). Acoustic structure and receiver response in domestic dogs, *Canis familiaris*. *Animal Behavior*, 39, 897-904.

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

Nakamura, G.V., Taraban, R., & Medin, D.L. (1993) Categorization by humans and machines. *The Psychology of Learning and Motivation*, Vol. 29. San Diego: Academic Press.

Nelson, D.A., & Marler, P. (1990). The perception of birdsong and an ecological concept of signal space. In W.C. Stebbins & M.A. Berkeley (Eds.), *Comparative perception: Complex signals*, Vol 2 (pp. 443-477). New York: John Wiley & Sons.

Nosofsky, R.M. (1986). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13, 87-108.

Ohl, F.W., Wetzel, W., Wagner, T., Rech, A., & Scheich, H. (1999). Bilateral ablation of auditory cortex in Mongolian gerbil affects discrimination of frequency modulated tones but not of pure tones. *Learning & Memory*, 6, 347-362.

Ohl, F.W., & Scheich, H., & Freeman, W.J. (2000a). Topographic analysis of epidural pure-tone-evoked potentials in gerbil auditory cortex. *Journal of Neurophysiology*, 83, 3123-3132.

Ohl, F.W., Scheich, H., & Freeman, W.J. (2000b). Spatial representation of frequency-modulated tones in gerbil auditory cortex revealed by epidural electrocorticography. *Journal of Physiology (Paris)*, 94, 549-554.

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

Ohl, F.W., Deliano, M., Scheich, H., & Freeman, W.J. (2003a). Early and late patterns of stimulus-related activity in auditory cortex of trained animals. *Biological Cybernetics*, 88, 374-379.

Ohl, F.W., Deliano, M., Scheich, H., & Freeman, W.J. (2003b). Analysis of evoked and emergent patterns of stimulus-related auditory cortical activity. *Reviews in the Neurosciences*, 14, 35-42.

Pearce, J.M. (1994). Discrimination and categorization. In N.J. Mackintosh (Ed.) *Animal learning and cognition*, (pp. 109-134). San Diego: Academic Press.

Phillips, D.P., Mendelson, J.R., Cynader, M.S., & Douglas, R.M. (1985). Response of single neurons in the cat auditory cortex to time-varying stimuli: Frequency-modulated tones of narrow excursion. *Experimental Brain Research*, 58, 443-454.

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

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4, 328-350.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7, 192-238.

Schouten, M.K.D, & De Long, L. (1999). Reduction, elimination, and levels: The case of the LTP-learning link. *Philosophical Psychology*, 12, 237-262

Wasserman, E.A., & Astley, S.L. (1994). A behavioural analysis of concepts: Its application to pigeons and children. *Psychology of Learning and Motivation*, 31, 73-132.

Wetzel, W., Wagner, T., Ohl, F.W., & Scheich, H. (1998). Categorical discrimination of direction in frequency-modulated tones by Mongolian gerbils. *Behavioral Brain Research*, 91, 29-39.