# The Role of Hierarchy in Learning to Categorize Images

**Reza Shahbazi (rs689@cornell.edu)**
Uris Hall, Dept. of Psychology, Cornell University
Ithaca, NY, 14853, USA

**David Field (djf3@cornell.edu)**
Uris Hall, Dept. of Psychology, Cornell University
Ithaca, NY, 14853, USA

**Shimon Edelman (se37@cornell.edu)**
Uris Hall, Dept. of Psychology, Cornell University
Ithaca, NY, 14853, USA

## Abstract

Converging evidence from anatomical studies (Maunsell, 1983) and functional analyses (Hubel & Wisesel, 1968) of the nervous system suggests that the feed-forward pathway of the mammalian perceptual system follows a largely hierarchic organization scheme. This may be because hierarchic structures are intrinsically more viable and thus more likely to evolve (Simon, 2002). But it may also be because objects in our environment have a hierarchic structure and the perceptual system has evolved to match it. We conducted a behavioral experiment to investigate the effect of the degree of hierarchy of the generative probabilistic structure in categorization. We generated one set of stimuli using a hierarchic underlying probability distribution, and another set according to a non-hierarchic one. Participants were instructed to categorize these images into one of the two possible categories a. Our results suggest that participants perform more accurately in the case of hierarchically structured stimuli.

**Keywords:** Hierarchy, Statistical Learning, Vision, Bayes, Probabilistic.

## Regarding hierarchies

The anatomy of the primate visual system suggests that the retinal input progresses through several stages of processing that form an approximate hierarchy. In the visual system, a large number of photoreceptors project to one ganglion cell, several of which converge onto a single LGN cell; then come the cortical areas V1, V2, IT, etc. (Kaiser & Hilgetag, 2010; Kandel, 2000; Modha & Singh, 2010).

The impression of hierarchy is further strengthened by evidence from functional analysis of the neuronal circuits. For instance, in V1 several simple cells send their axons to one complex cell whose preferred stimulus is constructed by the preferred stimuli of its input simple cells (Hubel & Wiesel, 1968). Moreover, starting from the retina and going up to higher cortical areas, the complexity of the features that each stage of this hierarchy responds best to increases (Gross, 1972).

There exist at least three different definitions of hierarchy in the literature. According to the most parsimonious of them, a hierarchy is any system of items where no item is superior to itself. Furthermore, there needs to be one hierarch, an item which is superior to all other items (Dawkins, 1976). This definition emphasizes that aspect of hierarchy that differentiates it from a heterarchy (McCulloch, 1945). According to McCulloch, heterarchy is a structure with a certain circularity. This circularity results in the possibility of members of the system being superior to themselves. Because of the paradoxes that it may engender, heterarchy is an unlikely structure to be observed in our everyday lives, hence the name (heterarchy is Greek for “under the governance of an alien”; Goldammer, 2003). Another definition of hierarchy comes from algebra, where hierarchies are defined in terms of partially ordered sets (posets; Lehmann, 1996). The third definition is the one advocated by Herbert Simon (1974 ), the pioneering figure of hierarchy theory. While the three definitions are not in disagreement with each other, the third one seems to be best suited for the present discussion.

According to Simon, a hierarchy is a nested collection of items where each item contains another set of subcollections. He uses the analogy of Chinese boxes, in which each box contains several smaller boxes while it is itself contained, together with other boxes, in a larger one. Graphically, this resembles the structure of a tree where vertices represent items and edges indicate containment. At least since the mid twentieth century, hierarchies have been believed to be the appropriate structure for the organization of complex systems in various domains including sociology, biology, computer science, and cognitive science (Simon, 1974; Hirtle, 1985; Holling, 2001).

In cognitive science, neuroanatomical data are one source of the evidence for the hierarchic structure of the visual system. Another line of evidence come from computational considerations. The problem of inferring the state of the environment from the sensory input is an ill posed problem (Chater, Tenenbaum, and Yuille, 2006; Edelman, 2008). The normative approach to this problem is to rely on the environmental statistics that have been acquired via past experience. A cognitive system that relies on the statistics of its environment to perform its tasks will soon run out of resources as the computational cost of keeping the joint statistics of the environmental variables grows exponentially in the number of variables that the system is keeping track of (an issue known as the curse of dimensionality; Bishop, 2006). By employing a hierarchic structure in recording the statistics, the system can bring the computational cost of the task under control. In addition to this computational advantage, hierarchic systems have been shown to be more stable and evolve faster than their alternatives (Simon, 2002).

While on the one hand it is inherently beneficial for systems to have a hierarchic structure, on the other hand, specifically in the case of perception, it is beneficial for a system to employ a hierarchic structure to represent its environment. Indeed, in the visual domain objects seem to present themselves to us in a hierarchic way. For example, a face is composed of two eyes, one nose, one mouth etc.; an eye is in turn composed of the iris, pupil, eyelashes etc. Is this hierarchy merely apparent, simply because of the hierarchic structure of our own perceptual system, or is it truly “out there”?.

In this paper we address this question indirectly, by evaluating the effect of the interaction between the probabilistic hierarchic structure that we build into a family of stimuli and the ability of human subjects to categorize those stimuli. In a series of related studies Aslin and colleagues have investigated learning of visual scenes in human subjects where higher level features are formed, in a hierarchical way, by chunking lower level features together. (e.g. Aslin et al., 2008). Here, we present participants with two sets of patterns composed of simple objects. In one of these sets, the scenes are drawn from a hierarchically structured probability distribution, while in the other one the dependencies are not strictly hierarchic. The subjects’ task is to categorize the patterns into one of the two possible categories. If hierarchies are an important aspect of the structure of the environmental systems, to which subjects are attuned, it should be more difficult for the participants to correctly categorize the non-hierarchic objects.

## The experiment

Participants were presented with images formed by twelve geometric shapes (figure 1) and were instructed to categorize them as either food or poison. Whether a certain image pattern is truly food or poison was initially unknown to participants, so that they needed to learn the diagnostic features by trial and error. Every time they responded “food” they were given auditory feedback (“correct” or “incorrect” tone). There was no feedback when they responded “poison.” Image patterns were sampled from probabilistic graphical models (a graphical representation of the joint distribution of the features in the image), specifically directed acyclic graphs (i.e. Bayes nets; Pearl, 2000; Bishop, 2008), designed to meet certain criteria. The Bayes nets had 12 visible nodes, comprising the image stimuli, and 10 hidden nodes (figure 2).

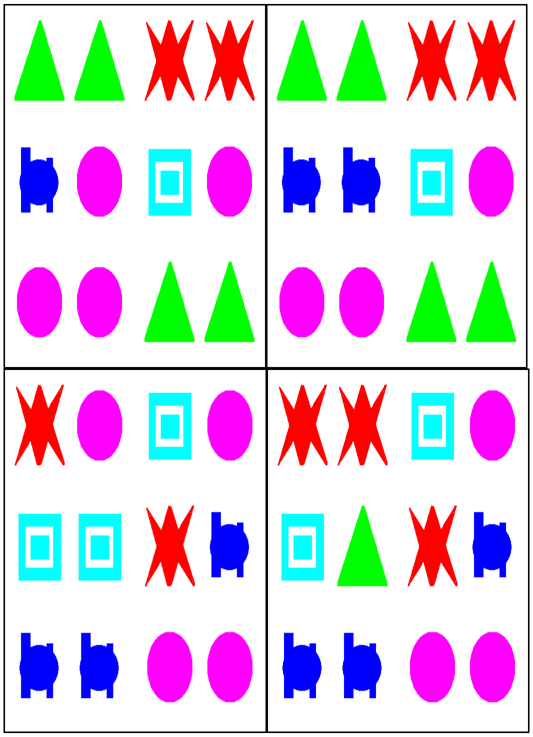

Figure 1: participants were instructed to categorize image stimuli into one of the two possible categories, “food” or “poison.” In one condition. The stimuli were generated according to a hierarchic structure. On the top two example images from this condition are presented which were designated as food items. In the other condition, images were generated according to a non-hierarchic structure. On the bottom two example food images from this condition are presented.

These hidden nodes represented the collection of contingencies upon which the nature of the image pattern (food or poison) relied. For example, one hidden node may denote the climate in which a certain fruit is grown, and another hidden node may denote the toxicity of the soil. In our experiment, the individual hidden nodes do not specifically stand for any such condition, rather the entire network of hidden nodes represents a typical network of causations, the end result of which makes the image a food or a poison. There were two sets of images: one sampled from a hierarchic Bayes net and the other from a non-hierarchic Bayes net. The non-hierarchic Bayes network was constructed in such a way that the image patterns sampled from its twelve visible nodes looked similar to the image patterns sampled from the hierarchic network. Note that in this setting, hierarchy is not an all or none property, and the non-hierarchic network still resembles, to some extent, a hierarchic structure (see concluding remarks for discussion).

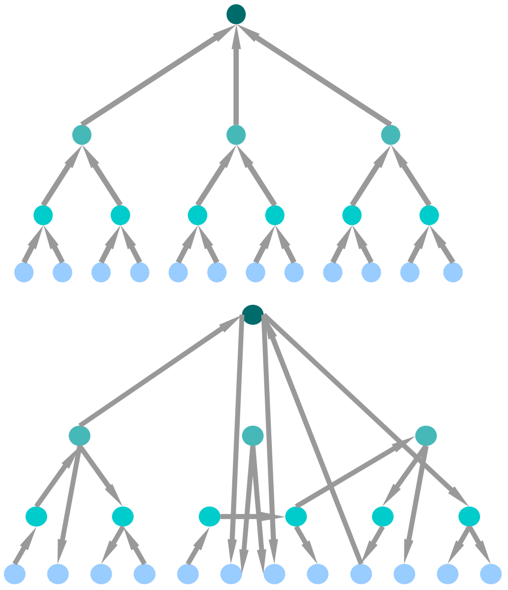

Figure 2: Graphical representation of the statistical dependencies in the hierarchic (top) and non-hierarchic (bottom) conditions.

## The Non-Hierarchic Case

For stimuli that are generated by a set of non-hierarchic causes, several factors may impair participants' performance. First, in the environment that participants are familiar with (the real world), causal structures are usually hierarchic. For instance, toxicity of the fruit is a feature formed by several lower level features (lower level merely in the hierarchic sense), such as the molecular structure of the soil, acidity of precipitation, ripeness (fruits that are too ripe are more prone to corruption), etc. We expected, therefore, that participants would try to utilize their existing hierarchic representation of the environment in learning the patterns, and that the mismatch between those representations and the causal structure behind the patterns would impair their performance. At the same time, non-hierarchic representations are more expensive to compute, and should add to the impairment of learning.

Furthermore, following the premise of statistical learning, participants are trying to learn the probability of a certain image pattern being associated with either food or poison: *Pr{F=food | I}*, where *F* denotes the nutrition content (i.e., food or poison) and *I* is the image pattern. The pattern consisted of twelve elements (the geometric shapes). Let us call them *e_i*. Therefore, *I=(e_1,...,e_12)*.

Keep in mind that even though the category of *I* is determined by nodes that are not directly observable, the effect of those hidden nodes must be accessible through the visible nodes, *I* itself. In fact, if the hidden nodes had no visible manifestation, learning the diagnostic features would be impossible. Therefore, observing that a subset of *I*, say, *(e_k,...,e_n)* has a particular value (e.g. 'star', 'star', 'triangle',...) counts as evidence in inferring the value of a certain hidden node. Ultimately it is the values of these hidden nodes that make the fruit food or poison.

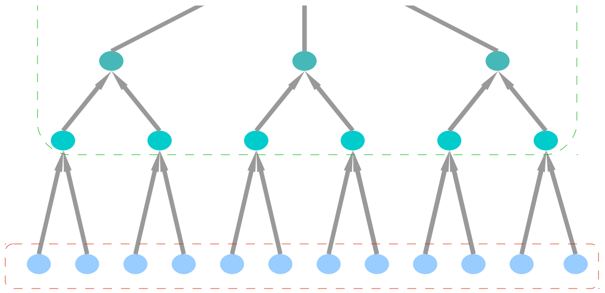

Figure 3: Images are sampled from the visible nodes (red dashed line) of the Directed Acyclic Graphs. Hidden nodes (green dashed line) represent the network of causes that determine whether the image is a food or a poison.

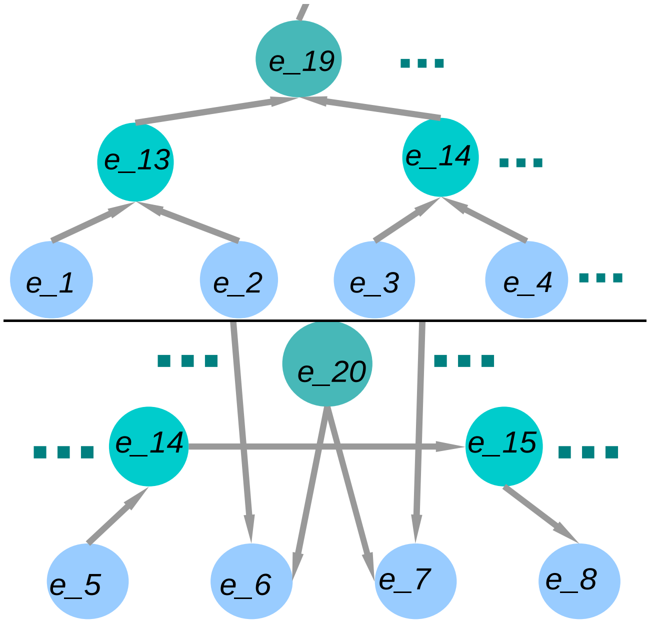

Figure 4: In the hierarchic condition proximal nodes' values are related locally (top) and their causal structure is more straightforward. In contrast, in the non-hierarchic condition (bottom) proximal nodes do not necessarily interact locally, and their causal structure is more complex.

In the hierarchic condition, *e_i* are related to each other in groups that interact locally (figure 4, top). For example, *(e_1,e_2)* are grouped together under the same hidden node; a hidden node from the second level of hierarchy, *e_13*. Similarly, two neighboring hidden nodes from the second level, *e_13* and *e_14*, are grouped together under a hidden node from the third level, *e_19*, and so on. In this situation, the values of the hidden nodes can be inferred in a straightforward manner by observing neighborhood clusters of the visible nodes. For instance, suppose participants have learned that the nutrition content of image patterns can be inferred based on the value of the first hidden node in the third level of hierarchy, *e_19*. The value of this particular node is reflected in the visible nodes *e_1* through *e_4*. Therefore, learning the required diagnostic feature amounts to learning the values of these four nodes. (Note that participants need not have explicit knowledge of the hierarchy. All they need to do is learn implicitly that certain configurations of *e_1* through *e_4* have a high correlation with poison or food).

In contrast, there is no such straightforward relationship in the non-hierarchic condition. First of all, visible nodes do not interact locally. For example, even though *e_5* through *e_8* are located close to each other, their features are contingent on hidden nodes which do not directly interact (figure 4, bottom). Furthermore, the statistical dependence may have a complicated structure: whereas in the hierarchic condition *e_5* through *e_8* ultimately depend on one hidden node, e*_19*, in the non-hierarchic condition *e_6* and *e_7* are governed by both *e_20* and *e_22*, while *e_8* depends on *e_5*, and *e_5* is conditionally independent of other nodes (i.e., its values do not depend on the values of the other nodes). The point is that even though there still exists a network of hidden causes that could in principle be used to infer the category of the stimuli, the more complicated structure of dependencies makes such inference more difficult to perform.

## Procedures

Eight participants (4 male and 4 female) took part in the experiment. Each participant performed both the hierarchic and the non-hierarchic conditions in a randomized order. Each condition consisted of 200 trials. It took each participant between fifteen to thirty minutes to complete the experiment. Images were presented on a computer screen using the Psychophysics tool box (Brainard, 1997) running under Matlab. In each trial, an image pattern was presented on the screen and participants had to respond by pressing either "Y", meaning they believed the stimulus was a food item, or "N" otherwise. There was no time constraint. The next stimulus appeared on the screen immediately after the participants' response. For each condition of the experiment, the participants initially started with 100 points – their remaining "life." For every "poison" item accepted, they lost 5 points; for every "food" item they gained 5 points. The last five "food" items that were correctly categorized were displayed at the bottom of the screen. Thus, feedback on the participants' choice was provided in the form of correct or incorrect only when they responded "Y".

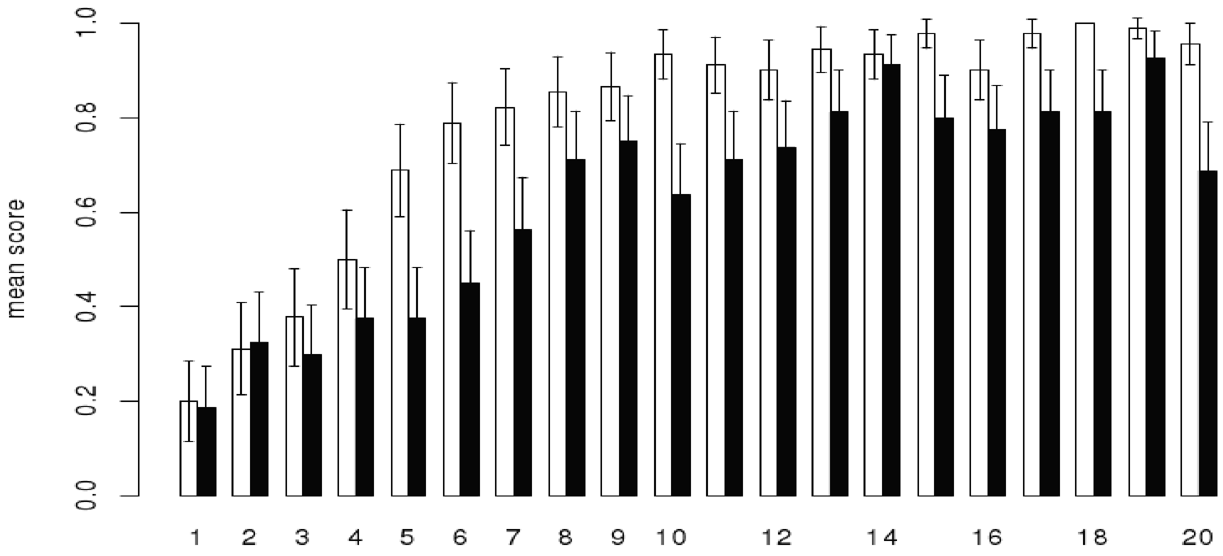

Figure 5: Comparison of performance in the hierarchic (white bars) versus non-hierarchic (black bars) over 200 trials. Error bars represent 95% confidence limits. Y axis shows mean accuracy of categorization over blocks of 10 trials.

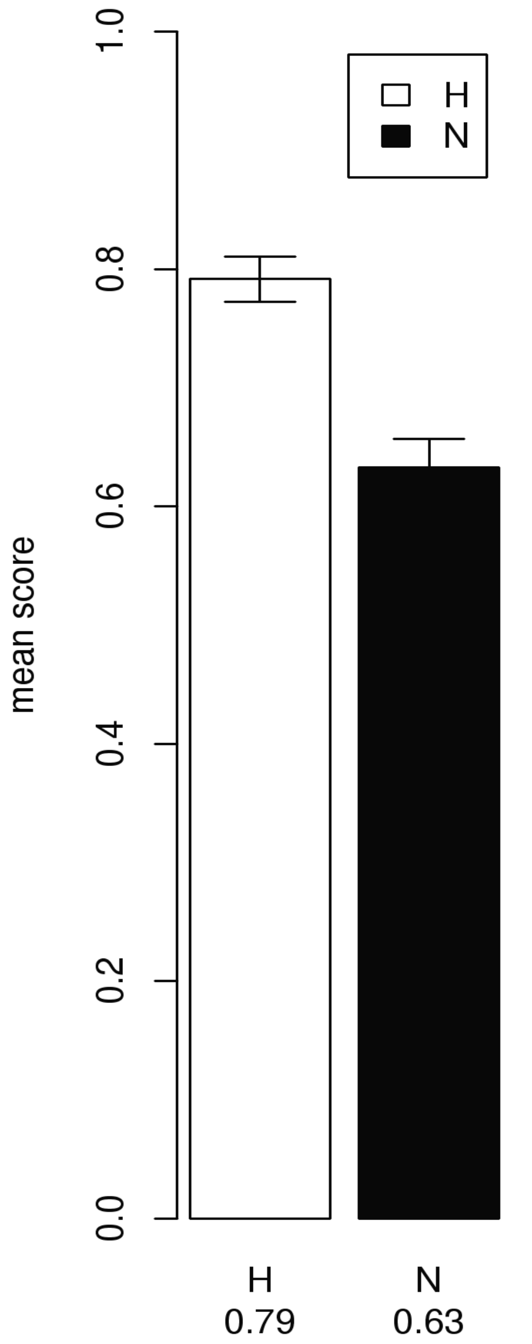

Figure 6: participants' performance measured as the mean percentage of correct classifications in each condition. H: Hierarchy, N: Non-Hierarchy.

### Results

Performance was measured as the percentage of correct classifications for each participant in each condition. On average, participants performed ~79% correct in the hierarchic condition compared to ~63% correct in the non-hierarchy condition (figure 6). This difference in performance is statistically significant as confirmed by the nonparametric Kruskal–Wallis rank sum test ($\chi^2$ = 104.91, df = 1, p < 2.2e-16). We also fit a linear mixed model to the data, to ensure that even when ll the random effects are considered jointly, significance is still reliable (Baayen, 2006), using the lmer procedure (Bates, 2005). A binomial logit-link linear mixed model fit to the scores yielded a significant effect of condition (z = 9.85 p < 2.2e-16). To explore the effect of gradual learning, we added trial number (in increments of 10) as an independent variable to the linear mixed model. In this analysis, the main effect of condition became n.s., the effect of trial number and the interaction between trial number and condition were both highly significant (z = 17.96, p < 2e-16, and z = 7.267, p < 3.67e-13, respectively; see figure 5).

## Concluding Remarks

There are several ways in which a structure can differ from a hierarchy. For example, links can skip levels, or the direction of the causation can be reversed. Consequently, further experiments are required to pin down the effect of each of them. Furthermore, the distinction between a hierarchy and a non-hierarchy is not all or none; rather it is a graded property, with perfect hierarchy at one extreme and heterarchy at the other extreme. We have been unable, however, to find a standard measure of the degree of hierarchicality in the existing literature. Developing and motivating such a measure is a topic for future work.

Another issue for future research is the possibility that subjects performed worse in the non-hierarchic condition of our experiment because the patterns in that condition were more complex. We plan to use the information entropy (Shannon, 1949) of the two graphs, as well as other measures of pattern generator complexity, in investigating this possibility. In the present study, we controlled for pattern complexity at the level of the leaves of the graph, by using stimuli that have the same appearance in both conditions.

## References

Baayen, R. H.(2006). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press, New York

Bates, D. (2005). Fitting linear mixed models in R. *R. News*5:27-30

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision* 10:433-436.

Chater, N. & Tenenbaum, J. B.,& Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences.* 10:287-291

Dawkins, R. & Bateson, P. & Gordon, P. (1976) *Growing points in ethology*. Cambridge University Press, New York

Edelman, S.(2008). *Computing the mind*, Oxford University Press, New York

Goldammer, E. & Newbury P. J. (2003). Hierarchy and Heterarchy. *Vordenker.* Retrieved from www.vordenker.de/heterarchy/a_heterarchy-e.pdf

Gross, C. G., & Rocha-Miranda, C. E. & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *J. Neurophysiol.* 35, 96–111.

Kaiser M., Hilgetag C. C., Kötter R. (2010). Hierarchy and dynamics of neural networks. *Front. Neuroinform.* 4112. doi: 10.3389/fninf.2010.00112.

Hirtle, S. C. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition.* Psychonomic Society

Holling, C.S., (2001). Understanding the complexity of economic, ecological, and social systems. *Ecosystems* 4:390–405

Hubel, D.H. & Wiesel, T. N., (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol,* 195 (1) 215-243

Kaehr, R., & Goldammer, E. (1989) Poly-contextural modelling of heterarchies in brain functions, *Models of Brain Functions*

Kandel, E. R., & Jessell, T. M., & Sanes, J. R. (2000), *Principles of Neural Science.* McGraw-Hill, New York

Lehmann, F. (1996). Big Posets of Participatings and Thematic Roles. *knowledge representation as interlingua —4th International Conference on Conceptual Structures*

Maunsell, J.H., van Essen, D.C . (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J Neurosci* 3(12):2563-2586.

McCulloch, W. (1945). A Heterarchy of Values Determined by the Topology of Nervous Nets. *Bulletin of Mathematical Biophysics*

Modha, D., Singh, R. (2010). Network architecture of the long-distance pathways in the macaque brain

. Orban, G., Fiser, J., Aslin, R. N., and Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105, 2745-2750. vol. 107 no. 3013485-13490

Orban, G., Fiser, J., Aslin, R. N., and Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105, 2745-2750, 105, 2745-2750

Pearl, J. (2000). *Causality*. Cambridge University Press, New York

Simon, H. A. (1974). *Hierarchy theory: the challenge of complex systems*. George Braziller, New York

Simon, H. A. (2002). Near decomposability and the speed of evolution. *ICC.* 11(3): 587-599.

Shannon, C.E. & Weaver, W., (1949). *The mathematical theory of information*. , University of Illinois Press, Chicago.