# Rules and associations : hidden Markov models and neural networks in the psychology of learning

Visser, I.

# 1 Cognitive rules

## 1.1 Introduction

Rule-guided behavior is central to many psychological theories. Cognitivism, the theoretical framework supporting rule-guided behavior, has dominated most areas of research in psychology since the demise of behaviorism in the 1950's. There seems to be general agreement that many behaviors *are* rule-guided, although the notion of rule-guided behavior is the subject of much debate. The behaviors associated with high level cognition, such as language production and understanding, problem solving, playing chess et cetera, are thought to be rule-guided. A set of rules that manipulate incoming information and produce output seems to provide and adequate description of such behavior. In the theoretical framework provided by such cognitive rules, a number of issues arise. First, we need to establish the nature of rules. Second, we need to know in what circumstances, or under which conditions, it is warranted to attribute rules to persons, or any other entity that might be suspected of following rules. Third, in order to further develop psychological knowledge, we need models that provide a viable implementation of rules.

There has been, and still is, plenty of debate about all these issues. In my masters thesis (Visser, 1996), I contrasted the Wittgensteinian and the Chomskyan conception of rules. I argued that the Chomskyan conception of rules lacks two important features, namely an embedding in behavior and an embedding in what Wittgenstein denotes a practice or institution (Wittgenstein, 1978). In particular, the problem with the Chomskyan conception is the looming infinite regression: symbols are analyzed or laid out in terms of other symbols which are further analyzed into still other symbols et cetera. At a certain point, this analysis has to stop, as Wittgenstein puts it:

> "If I have exhausted the justifications I have reached bedrock, and my spade is turned. Then I am inclined to say: 'This is simply what I do.' "
> (*Philosophical investigations,* § *217*)

Searle (1983) introduces the term 'Background' to describe the need for rules and symbols to be useful as such. The Background provides the necessary connection between a symbol and what it stands for in the world. Embedding of rules and symbolic structures in modes or patterns of behavior has been dubbed 'grounding' in recent literature (Harnad, 1990; Sun, 2000).

A necessary prerequisite for grounding rules and symbols is learning. Learning provides a means of tying mental states, i.e. representations, to behavior. Moreover, the learning process, as present in say language learning, also constitutes an embedding in a social practice in which language use is bound by certain rules. Hence, for models of cognition to be viable as models of rule-following behavior, learning is an essential feature.

In this thesis I study rule-following in three different guises. First, models of rule-following are considered. In particular the classical cognitive view on rule-following is contrasted with neural network models. I argue that these latter may also be interpreted as implementing rule-following behavior. Second, hidden Markov models are studied as statistical models of rule-following behavior that can be used to analyze both human behavior and neural network behavior. Third, implicit learning is studied as a task in which rule-following behavior is acquired by an associative learning process. In the remainder of this chapter I will address these three themes and provide an overview of this thesis. Before doing so however, some remarks on the nature of rules are in place.

Within the scope of this introduction, it is impossible to provide complete or even partial answers to the issues raised above about rules. The conception of rules that I am considering here is the classical conception of syntactic or formal rules (Chomsky, 1980; Fodor, 1981; Fodor and Pylyshyn, 1988). The origins of this conception are the notion of formal languages as it is put forward in the Chomsky hierarchy (Chomsky, 1959a) and the Turing machine (Turing, 1950/1990), which implements such languages. In this conception, rules are strictly syntactic or formal entities that manipulate (syntactic, formal) representations, i.e., the contents of a representation are irrelevant with respect to application of these rules. This is called the formality condition (Fodor, 1981). It is this conception of rules that Wittgenstein criticized. Apart from this formality condition of rules, another aspect of rules is important here. In general, cognitive rules need not be conscious. Hence, the conception of rules that is used here, is more liberal than is the case in many psychological applications. For instance, in developmental psychology children are classified as following certain rules in proportional reasoning, for example in solving the balance scale task. The rules that are used in models of this task are thought to be conscious rules, i.e. children are thought to consciously apply a rule in solving a given balance scale problem (see Jansen, 2001, for a discussion of children's strategies in solving the balance scale task and models for analyzing these).

The construction of models that implement rule-guided behavior is a major part of modeling in psychology. It may seem that the straightforward way of modeling rule-based behavior is by using the computational model of the Turing machine and the associated notion of formal languages. In fact, as the Turing machine is the paradigmatic machine that implements rule-following behavior, why should we not use it to model cognition? There are important problems in implementing cognition by means of the Turing machine model. In the Turing machine model, there is no natural way of grounding internal states, i.e., its representations. Next, there is the issue of learning that I discussed above. The Turing machine model has no natural place for learning as a process in which output is conditioned on feedback from its environment.

## 1.2 Rules in neural networks

The alternative to this classical model of cognition, which may be conceived as implementing syntactic rules and formal representations, is connectionism. Connectionism has gained much popularity over the past 15 years since the publication of the *PDP books* (Rummelhart and McClelland, 1986). Neural networks, more specifically feedforward or recurrent neural networks with supervised learning, provide models of cognition that naturally learn to represent a domain of knowledge on the basis of examples and feedback. However, neural networks are not without their own problems.

A first possible problem with neural networks concerns their representational or computational capabilities. In debates between connectionists and cognitivists, it has been put forward that neural networks do not have sufficient representational and computational resources to represent, say, human linguistic competence (Levelt, 1990). Fodor and Pylyshyn (1988) argue that even if neural networks prove to have these resources they would be 'merely' implementing classical rule-based structures. In chapter 2 of this thesis, mathematical results are discussed that provide insight into the computational capabilities of neural networks. The internal representations of such networks are compared to the mode of representation in classical models of cognition. This may provide answers to Fodor & Pylyshyn's arguments.

The second concern is whether neural networks implement rule-following. The way in which networks learn is often conceived of in terms of associative learning or conditioning. What then should be the criterion for deciding whether neural networks follow rules? At least, the tasks that neural networks learn or perform should be of the kind that we associate with rule-following behavior, e.g. language production or proportional reasoning. Next, we should be able to study the behavior and the internal representations of the network to answer the question whether they do indeed implement rules in some sense or whether they form different kinds of representations. These two points are addressed in chapter 3. Neural networks are trained to recognize languages and the representations that these networks form in the learning process are analyzed.

## 1.3 Psychometrics

One characteristic of rules that is pervasive in models of cognitive tasks, is that they may be unconscious or implicit. In the Chomskyan and Fodorian conception of rules that I sketched in the introduction, most rules are unconscious. For example, the rules that constitute knowledge of language, i.e. the rules that enable us to produce grammatically correct sentences, are unconscious (Chomsky, 1980). Also, the rules that are invoked in models of storing and retrieving memories, for example, are generally not open to conscious inspection. Indeed, it would be rather awkward if they were. Having to witness, within the confines of your own mind, say, the computation of the Bayesian a posteriori probabilities of two competing memory traces in a lexical decision task, would be, I imagine, a rather tedious experience. Hence, introspection is not always an option in trying to gain insight into cognitive processes. Most of the rules and representations that are used in psychological models and theories are unobservable entities. Even if they are open to conscious inspection, they would still be unobservable in the sense of objective observability. They would only be observable in the first person perspective, whereas a scientist requires a third person perspective. Hence, their existence can only be inferred from their observable manifestations in behavior. Psychometrics is explicitly involved in developing statistical models that relate observed variables to latent variables, which in turn stand for unobservable entities (see Borsboom et al., 2001, for a discussion of the relation between latent variables and the unobservable entities that they represent).

Many different latent variable models are in use in current work in psychology. Possibly the best known model is the common factor model which is invoked to order subjects on a latent trait such as intelligence, arithmetic skills or a personality trait. Markov models have been popular models in the area of learning and memory (Wickens, 1982). Latent or hidden Markov models form an extension of Markov models, which have a natural interpretation as models of rule-following. Hidden Markov models naturally allow this interpretation because they are equivalent to stochastic finite state automata, i.e. the canonical representation of simple grammars (Hopcroft et al., 2001). In chapters 4 and 5 of this thesis, statistical issues that arise in using hidden Markov models for psychological data are presented. Solving issues such as model selection and assessment of goodness-of-fit of latent variable models is crucial in their application to psychological phenomena.

## 1.4 Implicit learning

Implicit learning is an active field of research that concerns (the acquisition of) rule-following behavior. In this thesis implicit learning is studied as a form of associative learning that gives rise to rule-following behavior. In implicit learning grammatical structures are presented to subjects, which they have to reproduce or learn. During such experiments subjects are unaware that the material they have to learn is structured according to grammatical rules. Although we know from behavioral measures that subjects seem to grasp some of the structure underlying the stimuli, typically they remain unaware of what they have learnt. This is witnessed by subjects who say they 'felt they were responding appropriately' but can not provide the reason for responding in a particular way (cf. the quote from Wittgenstein in the first section).

Some researchers claim that the knowledge that subjects acquire in implicit learning is abstract, rule-based knowledge (Reber, 1993). On the contrary, others claim that such knowledge is merely a set of statistical associations shaped by the order of the stimuli (Cleeremans and Jimenez, 1998). Implicit learning described in this latter way is more akin to association learning or conditioning, and hence very different from rule-based knowledge. This difference in opinion expresses itself in differences in the models that are proposed to describe the knowledge acquired in implicit learning. Cleeremans and McClelland (1991) propose a neural network model that satisfies the statistical constraints inherent in the sequences of stimuli that are typically used in implicit learning. Using this model, they can account for about 80 % of the variance of reaction times in an implicit learning task. On the other hand, there are accounts in terms of chunk learning or rule induction (see Shanks and St-John, 1994, for a review of different accounts of implicit learning).

In studying implicit learning behavior, the same question arises as in studying neural network behavior: which criteria should be used to establish whether certain behaviors are based on rule-like representations, or whether they are merely statistical associations? The dividing lines between these views of implicit learning may not be as sharp as these authors claim. In particular, in the context of implicit learning it proves to be hard to differentiate between models in terms of their predictions. Wittgenstein (1978) argues that one of the ways of grounding rule-following behavior is a process like conditioning or associative learning. As I have stated before, the neural network model of implicit learning is usually viewed as a model of associative learning. Upon this interpretation, the neural network model is not in conflict with the rule-induction model but rather a detailed extension of it. In chapters 6 and 7 of this thesis, direct and indirect measures of sequence knowledge are compared in two implicit learning experiments in order to gain insight into the nature of the knowledge that subjects acquire in such experiments.

## 1.5 Overview of the chapters

In chapter 2, I discuss mathematical results relating to the computational capabilities of neural networks. Three major results are discussed that pertain to what can and what can not be represented in neural networks. The nature of representations in neural networks is said to be subsymbolic (Smolensky, 1988). I analyze neural network representations and compare them with the canonical representations of formal languages in the Chomsky hierarchy (Hopcroft et al., 2001). In so doing, the nature of subsymbolic representations is clarified.

I use hidden Markov models to analyze neural network behavior in chapter 3. This method of analyzing neural network behavior is contrasted with several other methods of gaining insight to the representations that neural networks build during learning. In particular, I analyze a neural network that has been trained to recognize the language of a finite state automaton. An empirical approximation of the entropy is used to compare fitted hidden Markov models to the language that the neural network was trained to recognize.

Confidence intervals are crucial in assessing whether a fitted model is adequate for a given data set. When fitting hidden Markov models to long timeseries, as is the case with implicit learning data, the standard method of estimating confidence intervals or standard errors of parameters on the basis of the Hessian fails due to computational problems. Hence, in such a case alternative methods are necessary to compute confidence intervals. In chapter 4, I compare three alternative methods of computing approximate confidence intervals.

In fitting hidden Markov models to psychological data a number of problems arise. First, model selection criteria are not readily available to compare fitted models. Second, goodness-of-fit measures are required to verify whether fitted models are adequate. Third, there is no general method for fitting models subject to equality constraints on the parameters, which is desirable in some applications. Possible solutions to these problems are discussed in chapter 5. I also present two applications of fitting hidden Markov models to psychological data in which the proposed solutions are applied. One data set is from a concept identification task and the other from an implicit learning experiment, which is further detailed in chapter 7.

In chapters 6 and 7, two implicit learning experiments are presented in which different measures of knowledge are compared. In chapter 6, performance on a prediction task is compared with performance on a reaction time task. In chapter 7, I compare performance on a free generation task with reaction time performance. In this chapter, a novel way of analyzing free generation with hidden Markov models is introduced. Hidden Markov models are used to characterize and quantify subjects' knowledge in terms of (stochastic) rules. This quantification of acquired knowledge allows for precise comparisons with performance on the reaction time task.

Finally, in chapter 8, I address the issues raised in this introduction, provide a summary of the main results, and sketch possibilities for future research. The chapters 2 through 7 of this thesis were all written as journal articles and have been published or are submitted. As a consequence, there is some redundancy in the text of this thesis, notably in the parts on hidden Markov models which are defined or described in four chapters. However, in all four chapter, their function is different and the description of the model is adapted to that function. Hopefully, the resulting emphasis on hidden Markov models will help to uncover their usefulness in psychological research.