Cognitive biases in neural networks - a study in reasoning under uncertainty

Ilkka Karanta, Timo Honkela, and Petri Koistinen

In: Proceedings of STeP-94, Finnish Artificial Intelligence Conference, pages 26–31, 1994.

Abstract

Although practical neural networks have not been designed mainly for imitating functions of the human brain, comparing the two might shed new light to the principles governing the behavior of both. In this study, we compare the behavior neural networks to that of the human brain with emphasis on cognitive biases in uncertain reasoning, and propose suitable experimental designs that could be used to reveal such biases. We also discuss how some aspects of the functioning of the human brain at a systems level could be simulated with neural networks.

Keywords

Human judgment, human information processing, cognitive processes, background knowledge, probabilistic reasoning, representativeness heuristic, availability heuristics, generalization capacity, anchoring, illusory correlation, simulating human organizations

References

W. Finnoff, F. Hergert and H.G. Zimmermann. Improving model selection by nonconvergent methods. *Neural Networks*, 6:771-783, 1993.

C.L. Giles, G.M. Kuhn and R.J. Williams. Dynamic recurrent neural networks: theory and application. *IEEE Transactions on Neural Networks*, 5(2):153-156, 1994.

R. Hink and D. Woods. How humans process uncertain knowledge: an introduction for knowledge engineers. *AI Magazine*, pages 41-53, 1987.

G.E. Hinton and D. van Camp. Keeping neural network simple by minimizing the description length of the weights. Computer Science Department, University of Toronto, 1993.

L. Holmström and P. Koistinen. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24-38, 1992.

D. Kahneman, P. Slovic and A. Tversky, eds. *Judgment under uncertainty: Heuristics and biases*. Cambrige University Press, New York, USA, 1982.

D. Kahneman and A. Tversky. On the psychology of prediction. In *Psychological Review*, pages 237-251.

T. Kohonen. *Self-organization and associative memory*. Springer-Verlag, second edition, 1998.

J. Sjöberg and L. Ljung. Overtraining, regularization, and searching for minimum in neural networks. Technical report, Linköping University, 1991.

P. Slovic and S. Lichtenstein. Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6:649-774, 1971.

S.E. Taylor. The availability bias in social perception and interaction. In Kahneman et al. *Judgment under uncertainty: Heuristics and biases,* chapter 13, pages 190-200.

A. Tversky and D. Kahneman. Judgment under uncertainty: heuristics and biases. In *Judgment under uncertainty: Heuristics and biases,* pages 1124-1131.

A.S. Weigend and N.A. Gerschenfeld, eds. *Time series prediction: forecasting the future and understanding the past*, volume XV of *Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis*, Reading, MA, 1994.

Cognitive biases in neural networks - a study in reasoning under uncertainty

Ilkka Karanta, Timo Honkela, and Petri Koistinen

Abstract— **Although practical neural networks have not been designed mainly for imitating functions of the human brain, comparing the two might shed new light to the principles governing the behavior of both. In this study, we compare the behavior of neural networks to that of the human brain with emphasis on cognitive biases in uncertain reasoning, and propose suitable experimental designs that could be used to reveal such biases. We also discuss how some aspects of the functioning of the human brain at a systems level could be simulated with neural networks.**

I. Introduction

It is well-known that human information processing – perception, judgment and choice (action) – under uncertainty is suboptimal at best, and the literature on the subject is extensive (cf. [3]). Less attention has been paid to what extent this suboptimality is also a property of neural networks, and how it could be modelled with their aid.

In this paper, we concentrate on biases in judgment. Probability theory has more or less been accepted as a standard of rational judgment under uncertainty. Therefore, it is of interest to compare how well inferential conclusions from a neural network compare to those obtained by probabilistic reasoning, and how these conclusions compare with conclusions of humans in comparable tasks.

Seeing whether and under what circumstances neural networks make the same mistakes in reasoning under uncertainty as humans do is of special interest, since human cognitive biases in this respect are well documented [6], and therefore we have a set of observations of irrational behavior that we can try to observe in the behavior of neural networks.

There are several contexts where this kind of information could be useful:

- simulation of human behavior in modeling societal systems or in complex simulation systems. Here the purpose is to use neural networks to simulate human perception, judgment and decision making as parts of models describing, e.g., groups or organizations.
- formulating tests for complex neural network systems when the tasks that the network should accomplish are poorly defined. If tasks can be found in which the neural network under consideration performs suboptimally, conclusions can be drawn as to whether it is appropriate for solving some ill-defined group of problems.
- understanding the mechanisms of such biases in biological neural systems. Naturally, even at best only a metaphoric view of the mechanisms that produce biases in the human mind can be obtained. However, finding a neural network with a human cognitive bias can lead to increased understanding of information processing in neural-like structures. It might shed light on, e.g., whether cognitive biases stem from inborn properties of the human brain or are they rather a result of culture and individual development.

II. Cognitive biases in reasoning under uncertainty

Human judgment has been compared to probabilistic and statistical reasoning by many authors (for further references, see [6]), and several patterns of human reasoning have been discovered. In the following, we will review briefly some of the more prevalent findings. We will use the terminology of Tversky and Kahneman [12].

It seems that people tend to assess probabilities of an object (or event) belonging to a class (or process) on how well the object typifies the class. This kind of inference heuristic is called *representativeness heuristic* [12]. It is fallacious in the sense that relevant statistical information such as prior probabilities and sample sizes tend to be overlooked. This is the way it shows in actual psychological tests.

One form of the representativeness heuristic is insensitivity to prior probabilities of outcomes. A sim-

I. Karanta is with Helsinki University of Technology, Laboratory of Information Processing Science, Espoo, Finland. E-mail Ilkka.Karanta@hut.fi

T. Honkela is with VTT Information Technology, Espoo, Finland. E-mail Timo.Honkela@vtt.fi

P. Koistinen is with University of Helsinki, Rolf Nevanlinna Institute, Helsinki, Finland. E-mail Petri.Koistinen@helsinki.fi

ple example of this is observed when people are told some features of a person and then asked to evaluate whether he is a lawyer or an engineer [7]. Although the people were told beforehand that the number of engineers in the sample is much higher than the number of lawyers, people ignored this base-rate information.

In certain situations, people seem to assess the probability of an event by the ease that instances or occurrences can be brought to mind [12]. The easier it is to recall instances, the higher the associated probability. This kind of inference heuristic is called *availability heuristic*.

The availability heuristic has been observed in social settings, among others. For example, if in a workgroup there is only a single member of a minority (e.g. woman, black or handicapped), this solo individual is taken as a representative of his or her social group [11]. Accordingly, the evaluations that are made of his or her performance are often used to predict how well other members of that group would do if they were to come into the organization as well.

In many judgment tasks, people make initial estimates and then adjust them to yield the final answer. These estimates may be based on some external source, or they may reflect initial guesses or preliminary computing. The adjustments made to these are typically insufficient [10]. This phenomenon is called anchoring.

Some well-known examples of anchoring occur when subjects are asked to estimate whether a certain ratio is bigger or smaller than a given random number, and thereafter estimate the ratio starting from the random number. The random numbers have a marked effect on estimates. Anchoring also causes biases in assessment of conjunctive and disjunctive probabilities, and of subjective probability distributions.

It is remarkable that all these biases seem to be tied to generalization. Representativeness is connected to generalization with the help of classes; availability is connected to some kind of inductive reasoning where the instances that can be brought to mind act as induction examples; and anchoring can be seen as generalization of an initial quess or first observations. Indeed, certain kind of anchoring has been used to improve the generalization cabability of neural networks [1]. An interesting question – though not the subject matter of this paper – is how intimately the capability to generalize and cognitive biases are interconnected.

III. Issues and problems in modelling reasoning under uncertainty with neural networks

There are many problems in modelling human reasoning with neural networks. Some are connected to the way human brains process information in general, and some to the specific tasks at hand.

An issue in modelling reasoning with neural networks is whether the neural network should directly reflect some assumed model of information storage and processing in the human brain. We have chosen not to assume any specific model of processing but rather to use simple neural networks that nevertheless in principle could be supposed to accomplish the given tasks. It is therefore doubtable whether the "cognitive biases" that emerge can be associated with actual cognitive biases. We claim that at a metaphorical level this is the case.

A question connected with this is: can some high-level functions of the human brain be reduced to the actions of small portions of it, and how small are these portions? This is an issue in neurophysiology, psychology and philosophy. We leave the question untouched.

Human information processing is dynamical in nature. On the other hand, the networks we consider in the present paper are static (for some references on a class of dynamic neural networks called dynamic recurrent networks cf. [2]). One might think that some parts of the human mind, such as long-term memory, are in a sense static in nature (weights of the network change with time but there is no "internal dynamics"), but this is not the case with human judgment. The latter point is explicit in some anchoring experiments where an essential part of the experiment is that the subjects are given a time interval too short to do precise problem solving. An example of this is an experiment described in [12]. The subjects were given the task of rapidly estimating the product of numbers from one to eight. However, anchoring can be found also in experiments where time limits are not so stringent.

One of the central problems in modelling cognitive processes with neural networks is that of representation. For example, numerical values such as those that can be given as an input to a neural network most certainly aren't represented in the brain as a single synaptic activation. We touch this question later, and show how a simple but slightly more realistic distributed representation can be implemented.

One factor that makes modelling cognitive biases with simple neural networks more unrealistic is the effect of background knowledge. Human beings trying to reason under uncertainty invariably possess a wealth of information more or less relevant to the

case at hand. This information has accumulated through the course of life, and modelling it – let alone taking it into a neural network model – is a very difficult task. We have chosen to almost completely ignore the role of background knowledge.

A problem in trying to model human reasoning under uncertainty is that some researches on cognitive biases are based on probability-based inference. It is hard to make a neural network do this.

One way of seeing whether judgmental biases in probability assesment can be induced in neural networks is to construct networks that directly assess probabilities and then trying to modify these networks. Several methods have been proposed for finding probability distributions of variables with neural networks. The simplest rely on having two outputs: the mean and the error bar. If normality is assumed, this information suffices to determine the distribution. A slightly more sophisticated way of accomplishing the task is to distribute the activation over a set of output units, allowing one to predict the probability density [13]. A weakness in these kind of approaches is that they require calculation of the percentiles, variances or mean values for the training sets outside the actual neural networks. Still another approach is to use "noisy weights": the precisions of the weights are also encoded [4]. It would seem that these kind of approaches don't have much resemblance to the actual functioning of the brain.

A problem in measuring biases in probability thinking is that the whole concept of probability is rather artificial from human point of view, and inference based on it doesn't necessarily represent any natural thought processes. Therefore probability inference is likely to be a high-level process in which for example processing of language is highly involved. It would hence be surprising if a any simple neural network would emerge that both had some cognitive biases comparable to human ones, and at the same time bore even a crude analogy to the actual processing of probabilities in the brain. We have omitted the study of actual probabilities from this study.

IV. Theoretical results and experiments

In the following, we will present some theoretical considerations and experiments that shed some light on how to design neural networks and data that metaphorically produce some cognitive biases.

To clarify the discussion, we present some notation here. The inputs to a neural network at a moment form an r-vector $\mathbf{P} \stackrel{\text{def}}{=} (\mathbf{P_1}, \ldots, \mathbf{P_r}$, and the outputs form an s-vector $A \stackrel{\text{def}}{=} (A_1, \ldots, A_s$. For compactness, we denote the combined input-output $r+s$-vector by $\mathbf{x} \stackrel{\text{def}}{=} \binom{\mathbf{P}}{\mathbf{A}}$. We call this vector the observation vector.

A. Representativeness

We study the representativeness heuristic in the following problem setting: let there be two groups of observations; one could represent, e.g., cooks, and the other astronauts. The observation vectors consist of normally distributed variables that could represent, for example, IQ, and physical fitness. There are much fewer observations in class 1 (say, the class of astronauts) than in class 2 (the class of cooks). The idea is that although the mean ability of the cooks is lower in all the abovementioned respects, the number of cooks is so much bigger and the variance of their abilities so large that there are more instances of cooks than astronauts in every sufficiently large part of the space.

More formally, we denote the number of cooks by n_1 and the number of astronauts by n_2, and so $n_1 \gg n_2$.

Representativeness can in principle be modeled in at least two ways. One is simply to see how the neural network handles the situation described above. Intuitively, a neural network classifier is very sensitive to the few observations in class 1.

Another way of modelling representativeness is to generate more observations to the class that naturally has fewer observations so that eventually, both classes have the same number of observations. We generate these artificial inputs by adding a small noise vector to some actual input, and treating the vectors so obtained as new inputs. More precisely, a number n of vectors is generated that deviate from the "representative" input $\mathbf{x}$ by a vector $\mathbf{v}_i$,

$$\mathbf{x}'_i = \mathbf{x} + \mathbf{v}_i, \quad i = 1, \ldots, n \tag{1}$$

and $\mathbf{v}_i$ are normally distributed and independent, their components also being independent of each other. In accordance our requirement, we set $n = n_1 - n_2$.

The idea of adding noise to actual inputs to obtain new inputs has received a lot of attention lately for example in improving the generalization capacity of a neural network [5]. Adding noise indeed improves network generalization. This leads to the natural question of whether availability heuristic is intimately connected to human ability to generalize, and whether the mechanism of generalizing patterns indeed could be based on forming transformed patterns from the original pattern by adding noise and feeding these new patterns to the processing units.

It would seem that making artificial observations to cancel prior probabilities is not the way that information is processed in the human brain. Indeed,

the representativeness heuristic apparently implies that humans form generalizations as models of entities, and then use these models in judgment rather than raw data. However, finding an alternative way of inducing representativeness heuristic may shed new light on the nature of the phenomenon. On the other hand, neural network models should be formulated which in a certain sense make models of real-world objects without regard to frequency of their occurrence. This problem is connected to pattern recognition as it would seem to be a viable way to consider these models as classes of patterns.

We will make the concrete experiment is as follows. We define two properties (that could be interpreted as IQ and physical fitness index). First we generate a set of normally distributed "observations" for the astronauts with mean vector $\mu_1 = (\mu_{11}\mu_{12})'$ and covariance matrix

$$\Omega_1 = \begin{pmatrix} \omega_{11}^1 & \omega_{12}^1 \\ \omega_{21}^1 & \omega_{22}^1 \end{pmatrix}$$

In a similar manner, a set of observations is created for the "cooks". In choosing the parameters, we keep in mind that the classes should not be linearly separable, and that the number of "cooks" should be much higher than the number of "astronauts".

B. Availability

Availability can be thought of as feeding the more available inputs to the network more often than those that are not so available. A natural idea is that availability could be imitated in a neural network setting by feeding some of the inputs and outputs of the training set to the network over and over again while neglecting some other input-output pairs. This would correspond to making some observations more "available" than others. However, this doesn't help in neural networks since most traditional learning schemes feed the inputs to the network many times anyway.

Another way of inducing the availability heuristic is to generate extra inputs from those inputs that are wanted to be available. Let the set of observations that we want more available than others be in the set $Q \stackrel{\text{def}}{=} \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}\}$. Then, in a direct analogy of the previous section, we generate

$$\mathbf{x}'_{ij} = \mathbf{x}_i + \mathbf{v}_{ij}, \quad j = 1, \ldots, m_{q_i}, \qquad (2)$$

for each $\mathbf{x}_i \in Q$. Here we have denoted the number of observations to be generated from $\mathbf{x}_i$ by m_{q_i}. As before, $\mathbf{v}_{ij}$ are normally distributed and independent, their components also being independent of each other.

We treat the specific instance of the availability heuristic that was described in section II. We are considering a set G of people, say women, that have a number of different attributes denoted by the vector $\mathbf{x}_i$. Some of these attributes, say attributes belonging to the set $U \stackrel{\text{def}}{=} \{x_{j_1}, \ldots, x_{j_u}\}$ (where each x_j is a component of $\mathbf{x}$) have been measured for only a small portion of the whole set denoted by G_k. This kind of attribute could be, e.g., the success of a specific woman in a particular kind of job.

The problem with ordinary neural networks is that the number of outputs is fixed. Therefore we have to find a special method for treating the above problem.

In principle, we could model the availability heuristic in the following way: for each person with attributes $\mathbf{x}_l$ in $G \setminus G_k$ we generate the missing attributes by adding a small random number to the $\bar{\mathbf{x}} \in G_k$ that most closely resembles $\mathbf{x}_l$. This is a viable way of modelling, but it is doubtable whether this resembles the way in which the generalization takes place in the brain. One would rather assume that there exists some kind of a general model of, say, women that is updated to incorporate new attributes when some observations of these attributes become available. Even if no observations are available of some attributes, humans can use contextual information to infer some plausible values for such attributes.

We take therefore a different approach here. The approach is based on the notion that information is represented in the brain in a distributed manner. This means that each input or output shows itself as a certain kind of activity structure rather than a definite activity value at a definite node.

We note that the output vector is located in R^s, the s-dimensional Euclidean space. Our idea is to rotate the outputs by a rotation matrix O (matrix O is such that its determinant is 1, and its product with its transpose O' is $OO' = O'O = I_s$, the s-dimensional identity matrix). If we want the original output values we just simply rotate back by matrix O'. Now, the output value isn't located in any single output node but distribued among all the outputs by the rotation. Although the formation of the output is straightforward and rather elementary it is based on a transformation that wastes no information.

A more general transformation would be obtained by adding to the result some suitable constant vector $\mathbf{W}$.

Use of neural network proceeds as follows: in training, we feed the network with the transformed outputs. Then we feed the network with the inputs for which we want to know the outputs. The real outputs are obtained from the outputs provided by the network by inversely rotating them by O'.

Now we can try to treat the problem of missing attributes in the following way: we set the missing attributes to zero in the training set. Then we feed the rotated $\mathbf{x}$ for all $\mathbf{x} \in G$, and when we want the value of some missing attribute, we rotate the output of the corresponding input back and read the value of the attribute from its proper place.

The procedures described above are partially based on the thought that availability heuristic, generalization and classification are closely intertwined, and that availability heuristic is a result of some observations receiving more processing in the brain than others. In our view, these presuppositions seem reasonable, but we don't know of any psychological research supporting these claims.

C. Anchoring and illusory correlation

Anchoring is a situation where the first impression too greatly influences the result of judgment [12]. An analog to this in neural networks is that learning coefficients are too low by an order of magnitude. Let us think of the neural network as a mapping

$$\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{w}), \tag{3}$$

where $\mathbf{w}$ is the vector that contains the parameters of the net. If $\mathbf{w}$ is estimated from data by an algorithm of stochastic approximation (see, e.g., [8]) type by updating the $\mathbf{w}$ on each round by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + a(n)\mathbf{g}(\mathbf{x}, \mathbf{w}(n)), \tag{4}$$

then, if $a(n)$ is way too small, the network sticks near the initial guess $\mathbf{w}(0)$. That is, the first observation has an unproportional effect on the final estimates.

The situation that there is a time limit to solving some problem can be modelled by thinking that the learning has been prematurely stopped. Stopping training prematurely is a well-known method of avoiding "overtraining" in neural networks [9].

Both of these experiments bear the implicit assumption that problem solving can be approximated as learning or, more precisely, stochastic approximation. It remains to be seen whether this is the case.

In illusory correlation, judges overestimate correlation even in cases where none exists or where negative correlation is present. Illusory correlations rise and persist because of a tendency to discount or ignore disconfirming evidence. This can also be modelled with the help of low learning rates, but here the factor that determines the final estimates is not the first observation but rather an imagined observation that is fed to the network before the actual inputs and outputs.

V. Conclusions

We have pondered the possibility of inducing analogs of certain cognitive biases in neural networks, and presented experimental arrangements with which such analogs can be found. It seems that the major types of cognitive biases in judgment under uncertainty can be simulated by neural networks. Experiments on the subject are still going on, and therefore it is still too early to say how well neural networks can accomplish the task. Results of this kind of research can be used in simulating human organizations with nets of neural networks, formulating tests for neural networks that are meant for ill-defined tasks, and improving understanding of how these biases occur in humans.

References

[1] William Finnoff, Ferdinand Hergert, and Hans Georg Zimmermann. Improving model selection by nonconvergent methods. *Neural Networks*, 6:771–783, 1993.

[2] C. Lee Giles, Gary M. Kuhn, and Ronald J. Williams. Dynamic recurrent neural networks: theory and application. *IEEE Transactions on Neural Networks*, 5(2):153–156, March 1994.

[3] Robert Hink and David Woods. How humans process uncertain knowledge: an introduction for knowledge engineers. *AI Magazine*, pages 41–53, Fall 1987.

[4] G. E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. Preprint, Computer Science Department, University of Toronto, June 1993. Referred to in Weigend and Gershenfeld (1994).

[5] Lasse Holmström and Petri Koistinen. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24–38, January 1992.

[6] Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press, New York, USA, 1982.

[7] Daniel Kahneman and Amos Tversky. On the psychology of prediction. In *Psychological Review* [6], pages 237–251.

[8] Teuvo Kohonen. *Self-organization and associative memory.* Springer-Verlag, New York Berlin Heidelberg, second edition, 1988.

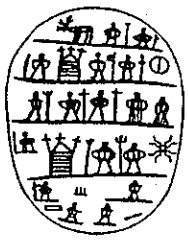
[9] J. Sjöberg and L. Ljung. Overtraining, regularization, and searching for minimum in neural networks. Technical Report LiTH-ISY-I-1297, Linköping Univ./Dept. of electrical engineering, S-581 83 Linköping, Sweden, 1991.

[10] Paul Slovic and Sarah Lichtenstein. Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6:649–744, 1971.

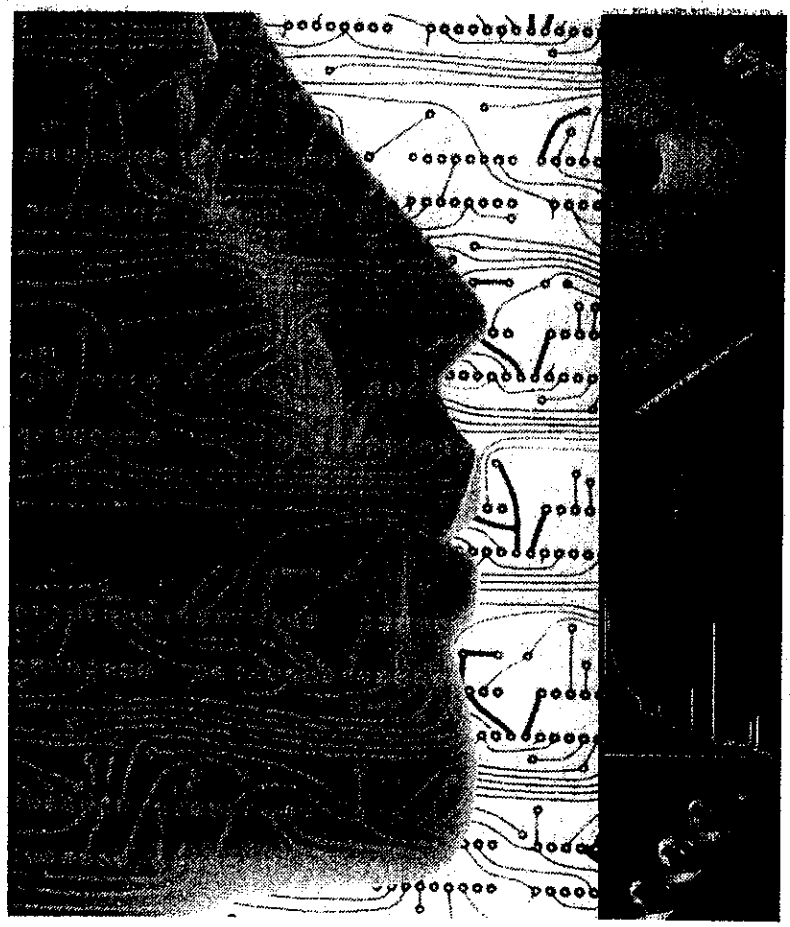
[11] Shelley E. Taylor. The availability bias in social perception and interaction. In Kahneman et al. [6], chapter 13, pages 190–200.

[12] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: heuristics and biases. In *Science* [6], pages 1124–1131.

[13] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time series prediction: forecasting the future and understanding the past*, volume XV of *Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis*, Reading, Massachusetts, 1994. Santa Fe Institute, Addison-Wesley.



Publications of the Finnish Artificial Intelligence Society (FAIS)



Multiple Paradigms for Artificial Intelligence

Proceedings of Contributed Session Papers

Conference on Artificial Intelligence Research in Finland
Suomen Tekoälytutkimuksen Päivät STeP–94
Turku Technology Center, 29–31 August 1992, Turku

Edited by

Christer Carlsson, Timo Järvi and Tapio Reponen

Åbo Academy University — University of Turku
Turku School of Economics and Business Administration