A Connectionist Model of Person Perception and Stereotype Formation

Christophe L. Labiouse & Robert M. French
Department of Psychology
University of Liege, 4000 Liège, Belgium
{clabiouse, rfrench} @ulg.ac.be

Abstract

Connectionist modeling has begun to have an impact on research in social cognition. PDP models have been used to model a broad range of social psychological topics such as person perception, illusory correlations, cognitive dissonance, social categorization and stereotypes. Smith and DeCoster [28] recently proposed a recurrent connectionist model of person perception and stereotyping that accounts for a number of phenomena usually seen as contradictory or difficult to integrate into a single coherent conceptual framework. While their model is based on clearly defined and potentially far-reaching theoretical principles, it nonetheless suffers from certain shortcomings, among them, the use of misleading dependent measures and the incapacity of the network to develop its own internal representations. We propose an alternative connectionist model - an autoencoder - to overcome these limitations. In particular, the development of stereotypes within the context of this model will be discussed.

1. Introduction

Until recently, connectionist models have had only a marginal impact on research in social psychology. However, in the last five years, researchers have tried to account for certain well established phenomena in the social cognitive literature using connectionist models. Among these phenomena are illusory correlation [7], cognitive dissonance [26, 30], person perception [28], impression formation [16, 17], and causal attribution [22, 29]. Read & Miller [23] brought together these disparate models in a review book dedicated to connectionist models in social psychology. Like Smith [27], we believe that connectionist modeling in social psychology may lead to a major theoretical integration of our understanding of social behavior and cognition. After a long period of conflict between the 'social cognitive' and the 'social identity' approaches, a connectionist approach to certain areas of social psychology could shed light on our understanding of stereotyping, prejudice, discrimination and other intergroup processes.

Smith & DeCoster [28] recently proposed what is, to our knowledge, the only connectionist model of social perception and stereotyping. They use a recurrent network based on the McClelland & Rumelhart's model of learning and memory [20]. They use a nonlinear activation update, bounded real-valued activations, and the delta learning rule. While their model is based on clearly defined and potentially far-reaching theoretical principles, it nonetheless suffers from a number of shortcomings and has, we believe, several methodological problems. The most

important limitation of their model is that it suffers from a linearity constraint in pattern learning: their model can learn a set of patterns only if the external input to each unit can be predicted perfectly by a linear combination of the activations of all other units, across the entire set of patterns. Smith & DeCoster suggest adding a hidden layer, something that is done in the model we propose here. There are a number of secondary problems with their model.

First, to account for the fact that people are able to generalize from multiple presentations of the same pattern, they train the network on a single pattern which is repeated 200 times (supposedly a frequently encountered individual), plus 1000 patterns randomly picked from a normal distribution (these are supposed to be the general background knowledge encountered by the social perceiver). To test a potential generalization, they probe the network with a corrupted version of the repeated exemplar. This task is unrealistic because people are exposed to many frequently encountered individuals. The problem with learning a single, often repeated exemplar is that this produces a very large basin of attraction for this pattern. Any starting pattern would likely reach the only available attractor state.

Second, we have had difficulty in reproducing their results. Using the equations they used, it would appear that the network has trouble finding a stable state in the activation update process. Formal analyses have shown that recurrent networks, in spite of their interesting dynamical properties, can also exhibit chaotic or oscillatory trajectories and cycle attractors [15]. This raises the question of when to modify the weights if a stable activation state has not been reached.

In this paper, we show that a simple multi-layer connectionist model - an autoencoder - can account for many robust phenomena in the social psychological literature. The rationale of our simulations was to develop a single model capable of qualitatively accounting for a wide range of well-known phenomena rather than fine-tuning one model to precisely reproduce the results of a single experiment.

2. Target Phenomena for these Simulations

We will test our model on a number of uncontroversial data patterns that can be found in the major handbooks and reviews of social cognition [9, 11], namely:

- *Exemplar-based inference*: Numerous studies [1, 18] have shown that people can learn particular properties of specific individuals (friends, family members, etc). With this knowledge, people can make inferences, often unconsciously, about unobserved traits or characteristics of a newly encountered individual.
- *Group-based stereotyping*: People can acquire stereotypes through social learning (gossip, media) and direct exposure to group members. They extract regularities in the traits of the people encountered and can apply this knowledge to draw inferences about either unobserved or perceived features (perceptual change in order to confirm the stereotype) of a new individual [14].
- *Concurrent exemplar-based inference and group-based stereotyping*: Our intent here is to demonstrate that a simple autoassociative memory can account for these two ways of processing that are generally difficult to integrate into a single framework. Traditional models [3, 8] have trouble integrating these two processes without resorting to a number of ad hoc hypotheses.

- *Development and formation of stereotypes*: We show that our model can account for many aspects of the development and formation of stereotypical knowledge without recourse to other factors, such as motivation, attention, cognitive load or norms. The use of stereotypical knowledge is seen as an increasing function of experience. Therefore, in our model, stereotyping can be conceived as a functional property in order to reduce the complexity of the social environment.

Only exemplar-based inference and group-based stereotyping were target phenomena in Smith & DeCoster's simulations. Most importantly, this model provides a unified theoretical framework for a fairly large number of phenomena related to stereotyping and social perception and can make novel predictions that can be tested in a traditional experimental setting.

3. Specific Aspects of the Model

Because of humans' necessary interactions with their social environment, the human brain has evolved in a such a way that social perceivers are able to cope with the intrinsic complexity of the social world. From this interaction and evolution have emerged, among other cognitive abilities, efficient face recognition, cheater detection, and one's own ingroup recognition. These abilities certainly provided an adaptive advantage during our evolutionary past. As a result, human brains are now able to recall a large number of individuals and events that allow them to deal with complex social situations. One of the most effective means of achieving this is to segregate the world into categories. Even if social and natural categories do not share the same properties, it is likely that the cognitive mechanisms by which we acquire them are similar. In both cases, categories could be extracted by a perceiver through statistical learning of the regularities in the world. Clearly, this is not an adequate explanation of all human learning but a large part of what a human being learns is probably implicit and requires no explicit rules or teachers.

In order to model certain cognitive social phenomena, particularly those involved in preconscious perceptual stages of conceptual interpretation, we chose to use an autoencoder whose task is to autoassociate a pattern via a hidden layer. This layer acts as a bottleneck and yields compressed representations of patterns. This network can learn without rules by observing exemplars, can automatically generalize, and can store precise information with a high degree of accuracy. Furthermore, autoencoders, unlike recurrent networks, do not suffer from the intrinsic problem of non-convergence to an attractor state. They also have a hidden layer which allows the network to overcome the problem of linear separability of the patterns to be learned. This hidden layer allows the network to develop its own internal representations, which it is certainly an essential feature of the human memory.

The results below are based on the performance of a 10-8-10 feedforward network. Activation values were either –1, 0, or +1. The rationale behind the coding is as follows: +1 could be conceived as the presence of an attribute or a trait, –1 as the absence of a trait, 0 was used to mean "impossible to determine whether the trait is present or not" (i.e. 0 is a "don't know" state. This coding fits the logic of social

interactions because it is frequently impossible to say if a person has a trait or not. Each pattern presented to the network represents an individual that a social perceiver could encounter. We never explicitly present a group in itself. The basic rationale of the simulations depicted here was inspired by a study by Mareschal & French [19] on early infant categorization. We use the standard backpropagation learning rule with momentum. The learning rate was set to 0.0001 and the momentum to 0.9. A Fahlman correction of 0.1 was applied. Networks were trained for a maximum of 100 epochs or until a error criterion of 0.2 for all outputs for all patterns was reached. The particular details of each simulation are given below.

4. Simulations

4.1 Simulation 1: Exemplar-Based Influence

The goal of this simulation is to see if the network is able to store 4 different frequently encountered patterns. Moreover, we want to assess the network's ability to generalize to novel exemplars. Although this property is a relatively well known feature of this class of networks, we performed this simulation in order to reproduce the Smith & DeCoster's simulations scheme as closely as possible.

Method & Results

The network was always given the same four bit-strings of length 10. However, to introduce variability, any bit could be randomly set to zero (i.e. "don't know" state) with a probability of 0.1. Each "participant" saw an equal number of each bit-string. We simulated 10 participants. The order of presentation of these 4 exemplars was randomized for each run.

We tested the network's memory for these exemplars by presenting two kinds of patterns. First, we probed the network with each "pure" exemplar (i.e. no 0's) to see if it learned to autoassociate the exemplar. Second, we probed the network with degraded versions of each exemplar to see how it fills in the blanks. For each exemplar and for each run, we present 10 "2-bit", 10 "3-bit", and 10 "4-bit" corrupted versions (i.e., 2, 3, 4 bits were randomly set to 0). For each of these two procedures, we computed an error measure consisting of the discrepancy between the actual output and the pure exemplar. When we probe with the pure exemplar, we expect to see a decrease of the error compared to the error level before training. Moreover, when we probe with new patterns, close to the original patterns, we expect a slight increase in error but significantly below the initial error for unlearned patterns. In other words, well learned exemplar representations influence the way new patterns, close to the originally learned patterns, are perceived.

The network performs as expected. Figure 1 shows the mean initial error score, the mean error score (after training) for the pure exemplars, and the mean error score (after training) for the approximate versions. The error is averaged over the 4 exemplars.

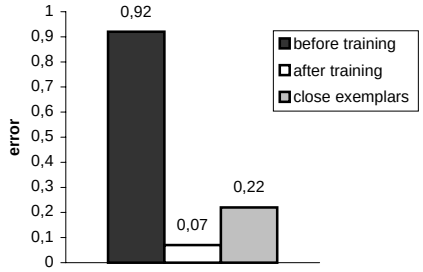

Figure 1: Mean error scores (Exemplar-based inference)

After learning, error is significantly lower, suggesting that the network has developed a reliable internal representation of the 4 exemplars. The generalization error rises slightly but stays well below the initial error, suggesting that the perception of similar exemplars is deeply influenced by the frequently encountered exemplars.

4.2 Simulation 2: Group-Based Influence

A stereotype is defined as a cognitive structure containing the social perceiver's knowledge, beliefs, and expectations about a group, learned through direct experience with individual group members and through social learning. Therefore, stereotypes affect inferences about newly encountered individuals [14] and these effects are often unintended and unconscious [5, 12]. Associations between a social category and a trait are most likely formed when they co-occur frequently and without too much variability [6].

Method & Results

We intended to show that the network is able to extract regularities in the presented patterns and can develop a prototypical representation of a group. Moreover, the network should be able to use this emergent knowledge to make inferences about new group members (i.e. patterns sharing some features with the prototypical group membership).

Instead of presenting the same bit-strings to the network, we presented variations of a prototype, ensuring that the network never encountered the prototypical group member. We continue to use the "don't know" state with a probability of 0.1. We simulate 2 groups, each consisting of 50 patterns. We first defined two bit-strings that will serve as stereotypes for the test phase. These two stereotypes were designed as follows: Five of the ten units were chosen as the "defining" features of the group because they are assumed to co-occur frequently. To introduce more variability, the probability that one of these units had the stereotypical feature was arbitrarily set to 0.8. The 5 other units were picked at random from randomly assigned values of 1 and –1. We simulated 10 participants. The order of presentation of the 100 patterns was randomized for each run.

We tested the network with the never-encountered stereotype to see if the network had extracted it from the repeated presentation of members. Second, we probed the network with incomplete versions of the stereotype to see how the network would fill in the blanks. For instance, does providing two stereotypical features allow the network to infer other stereotypical attributes ? For each stereotype and for each run, we presented 10 “1-bit”, 10 “2-bit”, and 10 “3-bit” corrupted versions (i.e., 1, 2, 3 bits set to 0). For each of these two procedures, we computed an error measure consisting of the discrepancy between the actual outputs and the desired outputs for the 5 “stereotypical” units.

Figure 2 shows the mean initial error score, the mean error score (after training) for the stereotype, and the mean error score (after training) for the previously unseen group members. Errors were averaged over both stereotypes.

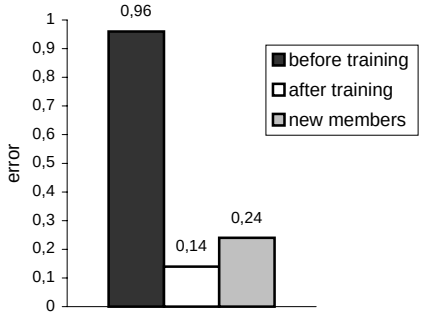

Figure 2: Mean error scores (group-based stereotyping)

After learning, the error is lower suggesting that the network has extracted regularities from the environment and has developed stereotypical knowledge. The generalization error is slightly higher but remains well below the initial error, suggesting that the network uses its newly acquired knowledge to infer stereotypical characteristics for new group members.

4.3 Simulation 3: Concurrent Exemplar- and Group-Based Influences

To investigate if the results of the two previous simulations were not artifacts, we test the network to see if it can exhibit both exemplar- and group-based processes simultaneously. A real social perceiver can simultaneously have a stereotype of a specific group and an accurate representation of exemplars which could be subtyped. The perception of newly presented individuals could be influenced either by the exemplar’s representation or by the stereotypical one. Does the present network have these properties ?

Method & Results

We build a set of patterns consisting of a single exemplar presented 50 times and 50 group members presented each once. The exact procedure was the same as in the previous simulations. We simulate 10 participants. After training, we probe the network both with patterns close to the exemplar or with “new group member”

patterns. We expect that, in both cases, errors will decrease compared to their initial state.

After training, the network exhibits both exemplar and group-based learning. Depending on the "person" (i.e., new pattern) encountered, the network is influenced both by frequently encountered exemplars and by emergent stereotypical patterns. Figure 3 shows the decrease in error, both for the "close-exemplar" and for the "stereotyped members" patterns.

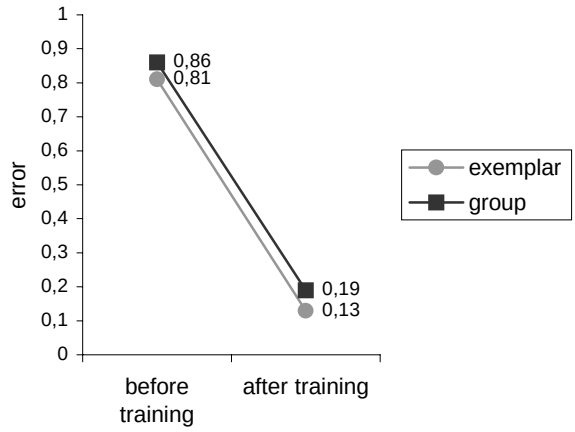

Figure 3: Mean error scores (concurrent exemplar-based and group-based stereotyping)

4.4 Simulation 4: Development of Stereotypical Knowledge

One important aspect of this model is that it can account for the development of mental representations of stereotypes. Sherman [24] noted that one of the most important factors influencing perceivers' reliance on exemplars or abstracted prototypes is the amount of experience perceivers have with the target to be judged. It is assumed that early encounters with a target will be of disproportionate importance because only a limited number of exemplars have been encountered from which to extract useful abstract knowledge. But as the number of encountered exemplars increases, a stereotypical representation can emerge, that then serves as the basis for subsequent inferences [25]. In this model, this property arises as a natural consequence of interacting with the complexity of the environment.

Method & Results

We took one of the stereotypes used in the second simulation and created a single exemplar from it. In the first phase, we used the same design as in the first simulation but with only this single exemplar instead of four. We then tested the network by computing two error measures: one with respect to the exemplar and the second with respect to the stereotype. The exemplar error was low and the stereotype error high. (See Figure 4). In the second phase, we presented the network

with a second set of patterns, which consisted of group members. This composition reflects the particular experimental design used by Sherman [24]. After this second phase, the same error measures were computed as before. We observe an increase in the exemplar error and a simultaneous decrease in the group-based error. These results are consistent with Sherman [24]. All results were averaged over ten runs and are shown in Figure 4.

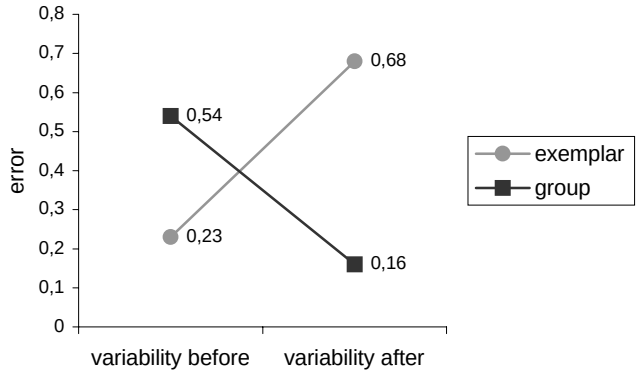

Figure 4: Mean error scores (Development of stereotypes)

5. Discussion

Compared to Smith & DeCoster's model, the autoencoder model is more powerful due to the presence of a hidden layer. This main improvement could be further tested in future studies (e.g. analysis of internal representations). In addition, autoencoders extract eigenvectors and are very close to PCA extractors. But they are more than PCA tools, insofar as they allow for learning, development and dynamical extrapolation of representations. However, although this model exhibits some interesting properties and can reproduce a number of important effects, it suffers from certain limitations. One of the most appealing properties of connectionist networks is their ability to do pattern completion, a property that is important in the simulations reported here. But the down side of this property is that the network will *always* fill in missing information whereas humans sometimes do not. This is certainly a drawback for current connectionist models of person perception. Sparse coding might be able to overcome this problem by producing a special "no recognition" state [4]. A second limitation of the model described here is catastrophic interference. Although learning is conceived somewhat differently in our simulations than in traditional cognitive psychology paradigms, the fact that catastrophic forgetting [21] can occur in these models is a real shortcoming for any model of human memory and cognition. This model, as any standard feedforward or Hopfield network models suffers from the "stability-plasticity" dilemma [13]. However, humans certainly do not suffer from catastrophic interference, especially with stereotypes. In fact, a stereotype is often particularly resistant to change. Modular computational architectures have been developed based on the brain's

hippocampal-neocortical division of labor to overcome this problem [2, 10]. Finally, by using an autoencoder, we lose the dynamic properties of an attractor network and, in particular, we lose the ability to study the evolution of attitudes or impression formation over time. Nevertheless we could undoubtedly overcome this limitation by adding recurrent links to the present model.

6. Conclusion

We have presented a simple model that captures certain properties of the early unconscious stages of social perception and stereotyping. The autoencoder, like humans, develops a relatively accurate representation based on single exemplars that can be automatically used to make inferences on newly encountered exemplars similar to those already encountered. Moreover, the network is able to reproduce these effects even in presence of variable inputs. This means that a stereotypical representation of a group can be extracted from repeated presentations of different members of the group. We do not claim that this is the only mechanism for stereotype formation and stereotyping but this statistical interpretation can arguably account for the early stages of these complex processes. We also show that exemplar-based inference and group-based stereotyping can be exhibited by a single autoencoder simulating the way humans store this type of knowledge. This network also offers a potential model of the development of stereotypical representations.

Acknowledgments

Christophe Labiouse is a Research Fellow of the National Fund of Scientific Research (Belgium). This work was supported in part by a Camille Hela grant awarded to C. Labiouse by the University of Liege, and by a research grant from the European Commission (HPRN-CT-1999-00065).

References

1. Andersen, S., & Cole, S. (1990). “Do I know you ?”: The role of significant others in general perception. *JPSP, 59*, 384-399.
2. Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Académie des Sciences de la vie*, 320, 989-997.
3. Brewer, M. (1988). A dual process model of impression formation. In T. Srull & R. Wyer (Eds.), *Advances in social cognition, Vol. 1* (pp. 1-36). Hillsdale, NJ: LEA.
4. Buhmann, J., Divko, R., & Schulten, K. (1989). Associative memory with high information content. *Physical Review A, 39*, 2689-2692.
5. Devine, P. (1989). Stereotypes and prejudice: Their automatic and controlled components. *JPSP, 56*, 5-18.
6. Dijksterhuis, A., & van Knippenberg, A. (1999). On the parameters of associative strength: Central tendency and variability as determinants of stereotype accessibility. *Personality and Social Psychology Bulletin, 25*, 527-536.
7. Fielder, K. (2000). Illusory correlations: A simple associative algorithm provides a convergent account of seemingly divergent phenomena. *Review of General Psychology, 4*, 25-58.

8. Fiske, S., & Neuberg, S. (1990). A continuum of impression formation, from category-based to individuating processes. *Advances in Exp. Social Psychology, 23*, 1-74.
9. Fiske, S., & Taylor, S. (1991). *Social cognition (2nd edition)*. New York: McGraw Hill.
10. French, R. (1997). Pseudo-recurrent connectionist networks: An approach to the "sensitivity–stability" dilemma. *Connection Science, 9*, 353-379.
11. Gilbert, D., Fiske, S., & Lindzey, G. (Eds.) (1998). *Handbook of social psychology (4th edition)*. Boston, MA: McGraw-Hill.
12. Greenwald, A. & Banaji, M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4-27.
13. Grossberg, S. (1982) *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Boston: Reidel
14. Hamilton, D., & Sherman, J. (1994). Stereotypes. In R. Wyer, & T. Srull (Eds.), *Handbook of social cognition* (2nd edition, Vol. 2, pp. 1-68). Hillsdale, NJ: LEA.
15. Hertz, J. (1995). Computing with attractors. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 230-234). Cambridge, MA: MIT Press.
16. Kashima, Y., Woolcock, J., & Kashima, E. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review, 107*, 914-942.
17. Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284-308.
18. Lewicki, P. (1985). Nonconscious biasing effects of single instances on subsequent judgments. *JPSP, 48*, 563-574.
19. Mareschal, D., & French, R. (1997). A connectionist account of interference effects in early infant memory and categorization. In *Proceedings of the 19th Annual Cognitive Science Society Conference* (pp. 484-489), Hillsdale, NJ: LEA.
20. McClelland, J., & Rumelhart, D. (1986). A distributed model of human learning and memory. In J. McClelland, & D. Rumelhart (Eds.). *Parallel Distributed Processing* (Vol. 2, pp. 170-215). Cambridge, MA: MIT Press.
21. Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308.
22. Read, S., & Montoya, J. (1999). An autoassociative model of causal reasoning and causal learning: reply to Van Overwalle's (1998) critique of Read and Marcus-Newhall (1993). *JPSP, 76*, 728-742.
23. Read, S., & Miller, L. (Eds.) (1998). *Connectionist models of social reasoning and social behavior*. Hillsdale, NJ: LEA.
24. Sherman, J. (1996). Development and mental representations of stereotypes. *JPSP, 70*, 1126-1141.
25. Sherman, J., & Klein, S. (1994). The development and representations of personality impressions. *JPSP, 67*, 972-983.
26. Shultz, T., & Lepper, M. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.
27. Smith, E. (1996). What do connectionism and social psychology offer each other ? *JPSP, 70*, 893-912.
28. Smith, E., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *JPSP, 74*, 21-35.
29. Van Overwalle, F. (1998). Causal explanation as constraint satisfaction: A critique and a feedforward connectionist alternative. *JPSP, 74*, 312-328.
30. Van Overwalle, F. (submitted). A feedforward connectionist model of cognitive dissonance: An alternative to Shultz and Lepper (1996).