Connection Science
Vol. 21, No. 4, December 2009, 359–377

Taylor & Francis
Taylor & Francis Group

Artificial neural networks whispering to the brain: nonlinear system attractors induce familiarity with never seen items

Serban C. Musca[a,b]*, Stéphane Rousset[a] and Bernard Ans[a]

[a]*Laboratoire de Psychologie et NeuroCognition, Université Pierre Mendès-France, CNRS UMR 5105, BP 47, 38040 Grenoble Cedex 9, France;* [b]*Laboratoire de Psychologie Sociale et Cognitive, Université Blaise Pascal, 34, Avenue Carnot, 63037 Clermont-Ferrand Cedex, France*

(*Received 14 November 2008; final version received 15 April 2009*)

Attractors of nonlinear neural systems are at the core of the memory self-refreshing mechanism of human memory models that suppose memories are dynamically maintained in a distributed network [Ans, B., and Rousset, S. (1997), 'Avoiding Catastrophic Forgetting by Coupling Two Reverberating Neural Networks' *Comptes Rendus de l'Académie des Sciences Paris, Life Sciences*, 320, 989–997; Ans, B., and Rousset, S. (2000), 'Neural Networks with a Self-Refreshing Memory: Knowledge Transfer in Sequential Learning Tasks Without Catastrophic Forgetting', *Connection Science*, 12, 1–19; Ans, B., Rousset, S., French, R.M., and Musca, S.C. (2002), 'Preventing Catastrophic Interference in Multiple-Sequence Learning Using Coupled Reverberating Elman Networks', in *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, eds. W.D. Gray and C.D. Schunn, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 71–76; Ans, B., Rousset, S., French, R.M., and Musca, S.C. (2004), 'Self-Refreshing Memory in Artificial Neural Networks: Learning Temporal Sequences Without Catastrophic Forgetting', *Connection Science*, 16, 71–99; Ans, B. (2004), 'Sequential Learning in Distributed Neural Networks Without Catastrophic Forgetting: A Single and Realistic Self-Refreshing Memory can do it', *Neural Information Processing-Letters and Reviews*, 4, 27–32]. Are humans able to learn never seen items from attractor patterns generated by a highly distributed artificial neural network? First, an opposition method was implemented to ensure that the attractors are not the items used to train the network, the source items: attractors were selected to be more similar (both at the exemplar and the centroïd level) to some control items than to the source items. In spite of this very severe selection, blank networks trained only on selected attractors performed better at test on the never seen source items than on the never seen control items. The results of two behavioural experiments using the opposition method show that humans exhibit more familiarity with the never seen source items than with the never seen control items, just as networks do. Thus, humans are sensitive to the particular type of information that allows distributed artificial neural networks to dynamically maintain their memory, and this information does not amount to the exemplars used to train the network that produced the attractors.

Keywords: human memory; distributed neural networks; memory self-refreshing; nonlinear system attractors; catastrophic forgetting; familiarity

The contribution of this paper consists in presenting a connectionist simulation and behavioural experiments that bring a new perspective on a kind of information that could be at the root of the basic functioning of human memory, namely distributed information. The general framework that

*Corresponding author. Email: serbancmusca@gmail.com

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540090903067493
http://www.informaworld.com

has inspired this research is the one where human memory is supposed to be well characterised as the end result of processes that are best approached through the dynamics of distributed nonlinear artificial neural networks (McClelland, McNaughton, and O'Reilly 1995).

However, while humans forget gradually, these distributed artificial networks forget catastrophically: newly learned information typically erases completely all previously learned information, which is quite a disconcerting property for models of long-term memory. This cognitively implausible phenomenon has been termed catastrophic interference – or catastrophic forgetting – (McCloskey and Cohen 1989; Ratcliff 1990). Various solutions have been proposed in order to avoid this major drawback of parallel distributed processing models (Hetherington and Seidenberg 1989; McCloskey and Cohen 1989; Kortge 1990; Ratcliff 1990; Lewandowsky 1991, 1994; Murre 1992; French 1992, 1994, 1997; McRae and Hetherington 1993; Lewandowsky and Li 1995; McClelland et al. 1995; Sharkey and Sharkey 1995; Robins 1995, 1996; Ans and Rousset 1997, 2000; Robins and McCallum 1998, 1999; French, Ans and Rousset 2001; Ans, Rousset, French and Musca 2002, 2004; Ans 2004; for reviews, see French 1999; Blackmon, Byrd, Cummins, Poirier and Roth 2005). In this paper, the solution involving a memory self-refreshing mechanism based on a reverberating process is the one that will be considered because previous work of ours has proven it to be efficient (Ans and Rousset 1997, 2000; Ans et al. 2002, 2004; Ans 2004), and it distinguishes itself from the other solutions as it is one of the only two (along with that of French 1997) that offer an all-distributed solution to the problem of catastrophic forgetting. As it will be exposed, it also has characteristics that make it a suitable method when addressing the topic of learning from distributed information in humans. It supposes a mechanism whereby a network's memories are dynamically maintained through self-generated nonlinear system attractors, which are the outcomes of random input activities reverberated many times within the neural network. This means that, thanks to the memory self-refreshing mechanism that allows for a pseudorehearsal within the network, information already learned by the network is not lost each time new information is learned.

The main objective of the present paper is to answer the following question: Are these attractor states only formal entities, or have they also a psychological significance? If the latter is true, then we expect humans to be sensitive to attractor patterns generated by an artificial distributed network. The attractors being closely linked to the very nature of distributed information and to the memory self-refreshing mechanism, let us first see these notions before exploring if the attractor patterns are, or are not, processed by humans.

Distributed information originates from the functioning mode of parallel and distributed nonlinear neural networks, in particular from that of a highly representative class of such networks, that of multi-layered networks trained by a gradient descent learning procedure – of which the most popular is the backpropagation learning algorithm. For concision sake, we hereafter call a network of this type a *GDN* (for gradient-descent trained network). Distributed nonlinear networks in general and GDNs in particular distance themselves from previous models of learning/memory by the fact that the training exemplars are no longer stored as such in the system but only contribute to shape its 'internal landscape'. Indeed, the memory of such a network gradually emerges through the processing of the training exemplars by the learning algorithm. As a result of this training, the weights of the connections between the processing units reach values that allow the network to perform correctly. The memory of a trained GDN can thus be conceived as the particular set of connection weights between its processing units. This distributed nature of information, essentially required within networks to achieve generalisation – a key property of these systems that makes them appropriate models of human cognition – seems to be incompatible with a low 'forgetting' level: because in such highly distributed systems knowledge representations about different learned items extensively share the same connection weights, when a new set of items is learned, the same connection weights, which were already adjusted for previously learned items, will be once more modified. As a result, learning of new information may completely abolish

memory of old information, resulting in the classical 'sensitivity–stability dilemma' (Hebb 1949) or 'stability–plasticity dilemma' (Grossberg 1987; Carpenter and Grossberg 1988), now termed catastrophic forgetting.

The neural network architecture proposed to overcome catastrophic interference (Ans and Rousset 1997) consists of two coupled GDN networks, NET1 and NET2, each of them having the same functioning and the common basic structure shown in Figure 1 – except that the number of hidden units can be different in the two networks. Within each network, an input layer is fully forward-connected to a hidden layer, itself fully forward-connected to a hetero-association layer. In contrast with classical feed-forward networks, the hidden layer is also fully backward-connected to the input layer, which means that the network is not only trained to associate each input to the correct hetero-associative output (e.g. associate each face with its name) but also to correctly reproduce each input. The backpropagation learning algorithm is used to associate input to target patterns. Typically, a set of input-target pairs (the to-be-learned training base) is repeatedly presented to the network. At each presentation of a pair, all the network's connection weights are differentially modified so as to minimise an error function based on the error between the output activation actually computed by the network in response to the input, and the target provided for each pair. Here, this error includes not only the usual error between the hetero-association layer activation and the hetero-associative target, but also the error between the computed output activation from hidden units to input units and the input pattern – i.e. this latter plays the role of an auto-associative target. Thus, the input layer is also an auto-association layer. It is noteworthy that this auto-associative part is always required in the basic network (for the reverberating process implementation, see below) even when a hetero-association is the focus of the task under study.

The goal of the dual architecture is to avoid catastrophic forgetting in the primary network, NET1, whose task is learning items that come from the external world (see Figure 2 for a diagram showing the flow of information between the two networks). This is achieved with the help of

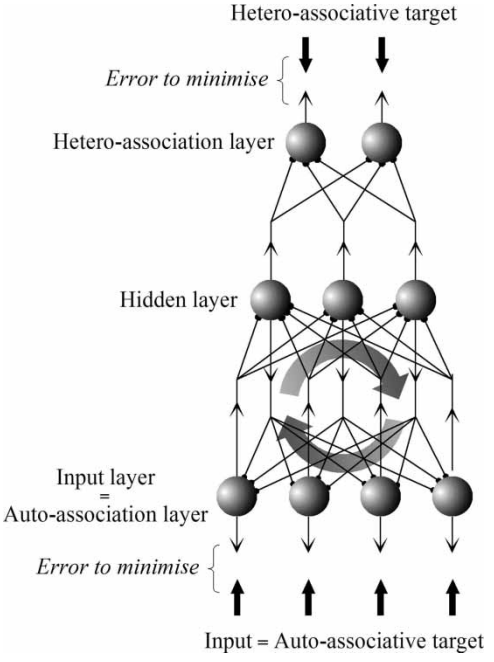

Figure 1. Example of an auto-hetero-associative distributed neural network. The more general architecture of Ans and Rousset's (1997) memory self-refreshing mechanism is made up of two such networks (see text for details).

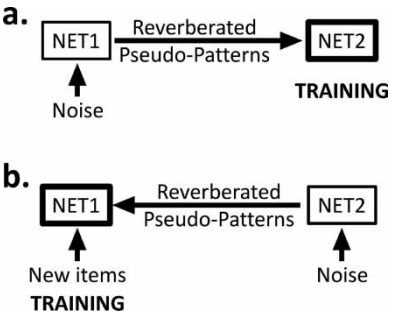

Figure 2. Flow of information between the two coupled hetero-associative GDN networks, NET1 and NET2: (a) Stage I: once trained, NET1 generates RPPs from random noise and NET 2 is trained on these RPPs; (b) Stage II: once NET2 trained, NET1 is trained on both new items and on RPPs generated from random noise in NET2. See main text for details.

the secondary network, NET2, which is not in contact with external events but only with NET1. For a simple explanation of the basic functioning of this dual architecture let us consider an initial state where NET1 has already completely learned a given set of external input-target pairs, and NET2 is still 'empty' (i.e. with random connection weights). Assume that the neural system then enters in a first processing phase, Stage I, where NET1 is no longer receptive to external events (i.e. it does not learn anymore, so its connection weights do no longer change) but is 'bombarded' over its input layer with random stimulations issuing from a noise generator. For each such stimulation, a first resulting output activation arises at the hetero- and auto-association layers by a mere propagation along the network's connections. This first auto-association layer activation is then re-injected to the hidden layer and results in new output activation at the hetero- and auto-association layers. This second auto-association layer activation is re-injected again in the hidden layer, hence recreating a third new network activation, and so on. This back and forth flow of activation between the hidden and input layers is termed 'reverberating' process. Ideally, this cycling process goes on until convergence to an attractor state of the neural network, but in practice an attractor can be approximated by the activation pattern produced after a fixed number of re-injections within the network's auto-associative part. The pattern obtained in this way, including an auto-associative activation and the corresponding hetero-associative activation, is called a reverberated pseudo-pattern (RPP). The two components of the current RPP from NET1 are then respectively transmitted to the corresponding NET2 layers for training. The first NET1 component plays the role of both an input and an auto-associative target for NET2, the second component playing the role of a hetero-associative target for NET2. A large set of RPPs generated in NET1, each from a different random stimulation, is used to train NET2. Thus, Stage I is intended to 'transport' learned information from NET1 to NET2.

Stage I is followed by a second procedure, Stage II. Stage II aims at allowing NET1 to learn new external items while not forgetting the old ones. During this stage, NET1 will learn concurrently new external items interleaved with RPPs generated this time in NET2, these RPPs reflecting NET1's previous knowledge. Indeed, the weight modifications required in NET1 to learn the new items are now not only constrained by these new items but also by NET1's previous knowledge, information that has been transferred to NET2 (through RPPs generated in NET1) during Stage I. In the course of learning, Stages I and II follow one another for each new occurring population of to-be-learned external events. This self-refreshing of the neural system memory by RPPs has proven to be an efficient way to overcome catastrophic forgetting in sequential learning tasks (Ans and Rousset 1997, 2000; Ans et al. 2002, 2004).

As already mentioned, NET2 does not receive information directly from the environment but only from NET1. Moreover, in order to avoid catastrophic forgetting, it is NET2 that provides NET1 with information on the knowledge already acquired by NET1 when this latter is learning

new information. These facts indicate that the memory self-refreshing mechanism is based on the continuous flow of network attractors back and forth between NET1 and NET2. Now, what information is there in the network attractors created in NET1 and used to train NET2? The answer to this question is of topical importance to this paper.

Let us assume that NET1 has learned a set of items and generates RPPs that are used to train NET2. Let us further suppose that RPPs are a blend, some of them being the actual items that were used to train NET1 and some of them being something else, attractors that are not items (non-item attractors, hereafter, NIAs). The previous question can now be rephrased: Are the attractors of this second type just noise, or do they convey information on the actual items, while not being these items? To check if this is the case is to verify whether information on a set of items can be efficiently conveyed by NIAs. After exposing a method that ensures the attractors that will be used are not the actual items that were used to train NET1 (under their initial form or as noisy, distorted versions of them), it will be checked whether information on a set of items can be conveyed through NIAs from a neural network to another neural network, and also from a neural network to humans.

When neural networks are concerned, after a blank network, NET2, learns NIAs generated in another network, NET1, there are measures that allow one to assess how well NET2 performs on the set of items NET1 had been trained on although NET2 had never been trained directly on these items. NET2 is only tested on these items. Transposing this to the situation where humans are concerned, it amounts to testing the memory for never seen items. Interestingly, there is a way to devise such a test: considering the influential dual-process model of recognition memory (Mandler 1980), the recollection process will not serve (since there is not such a thing as recollection of the encoding episode for items that had never been seen before), but the familiarity process can be used to test the memory for never seen items. This method entails, however, a slight complication because, unlike an absolute performance measure in neural networks, a measure of familiarity with some items cannot be used as an absolute dependent variable. It only makes sense in a comparison including a measure of familiarity with other items. That is, a relative measure of familiarity shall be used, and more concretely a comparison of the familiarity with the items of interests (i.e. those items that were used to train the network that generated the NIAs) to the familiarity with some other items of the same kind.

The results of the forthcoming connectionist simulation and behavioural experiments, taken together, will speak to the issue of the nature of information that is conveyed by NIAs and, more importantly, to the issue of the ability of the human cognitive system of making use of this kind of information. So the aim of this study is not to evaluate the whole neural network architecture previously proposed to overcome catastrophic interference but to assess the psychological validity of NIA-type information and its specificity compared with other kinds of information. Since NIAs take root in the very nature of distributed memory, the present study could more generally offer new pieces of evidence in favour of the distributed nature of human memory.

1. Creating the NIAs

This section presents the rationale and the method that will allow for a comparison of familiarity between two sets of items, on the one hand the items of interest (hereafter *source items*), that is, those items that were used to train the network that generated the NIAs, and on the other hand, the *control items*, that is, some other items of the same kind that are not used as training material. The prediction that we want to check is whether RPPs can convey information on the source items without being these items. In order to do this, it is crucial to ensure that the selected RPPs, i.e. NIAs, are not the source items, under their initial form or as noisy, distorted versions of them. The

method through which we achieve this here, in the context of a future comparison of performance on source and control items, consists in ensuring that NIAs are not only very different from the source items but – with regard to the similarity dimension – are even closer to the control items. Because of its extreme nature, we call this an *opposition method*. This reasoning naturally leads to the idea of selecting among the RPPs following some selection rules. These rules and their reason to be are briefly exposed here then detailed later on. A first rule ensures that at the exemplar level each of the NIAs is closer to a control item than to any of the source items. A second rule considers similarity at the level of the whole set of NIAs, and ensures that the 'mean NIA' (i.e. the centroïd of the set of retained NIAs) is closer to the control items than to the source items[1].

As pointed out before, the only way to test the memory for never seen items is through a test of familiarity, and more precisely by comparing familiarity with the never seen source items to the familiarity with the never seen control items. Thus, participants in the behavioural experiments will first be exposed to the set of selected NIAs (displayed as visual patterns) then will undergo a familiarity test on the never seen source and control items. In neural network terms, familiarity is restricted to the auto-associative part of the general model presented in Figure 1. Indeed, from a neural network viewpoint a familiarity task does not require that the network learns hetero-associative targets but only to estimate whether the stimulus at hand (that is, the input) has previously been processed or not. This latter task is by definition the task of the auto-associative part of the model, also termed for this reason auto-encoder. Thus, with respect to the architecture in Figure 1, the hetero-association layer is not required, so the neural networks that will be used will only comprise an auto-association layer and a hidden layer, and their task will be, after training, to reconstruct over the input layer the items that are presented to it.

1.1. *Stimuli*

The stimuli to be used as source items and as control items both come from a single and homogenous family of items. The items are matrices (cf. Figure 3) constructed as follows. Starting from the centre of a 19×19 black grid, the following procedure was applied 20 times: A direction (up, down, left or right) was randomly chosen and two squares in that direction were turned white, then the last square served as starting point for the procedure on the next step. Any resulting pattern wider or higher than 13 squares was discarded, the remaining were re-centred on a 13×13 grid until 106 different and meaningless items were available. The set of 106 items was *randomly* divided into two subsets of items, Lists A and B, for the Simulation and Experiment 2 (cf. Figure 3). For Experiment 1, the set was *randomly* divided into three subsets of items. For

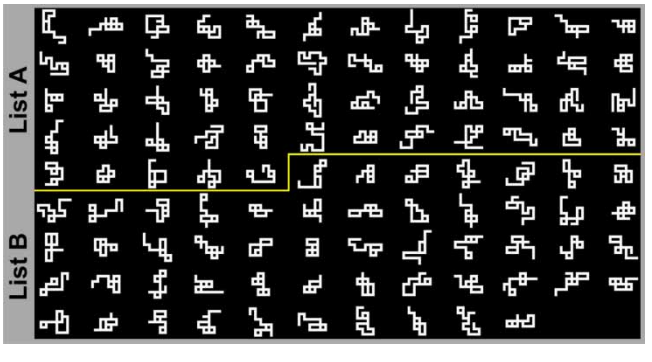

Figure 3. The full set of items used in the simulations and the behavioural experiments (and the random division into Lists A and B for Simulation and Experiment 2).

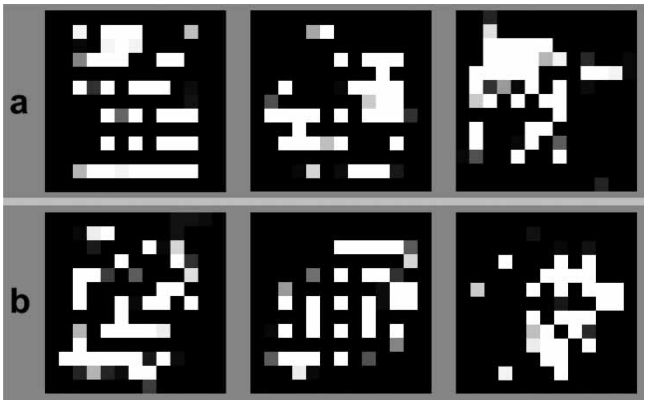

Figure 4. Examples of selected RPPs: (a) three NIAs from NIA_A (generated by a network trained on List A), and (b) three NIAs from NIA_B (generated by a network trained on List B), corresponding to a source list counterbalancing (see text for details).

the neural networks each item was coded as a vector of length 169, with black squares coded 0 and white squares 1.

1.2. *RPPs and NIAs*

To generate the RPPs, a backpropagation auto-associator with a 169-unit auto-association layer and a 16-unit hidden layer was used. All units had a sigmoid activation function. The learning rate was of 0.01 and the momentum of 0.7, which is a couple of parameter values in the range typically used in order to obtain good learning through a smooth-gradient descent. The usual backpropagation learning algorithm that minimises the cross-entropy cost function (Hinton 1989) was used.

Initialised with random connection weights uniformly sampled between -0.5 and 0.5, the network was trained on the source items until the error on each component of each training item was less than 0.01. Once this criterion reached, the connection weights were not changed any more. The RPPs were then generated in the following way.

A random 169-component pattern was fed to the input layer that produced, after a mere propagation along network's connections, a first output activation in the auto-association layer. This first resulting activation was then re-injected to the hidden layer and resulted in new output activation at the auto-association layer. This secondly produced activation was re-injected again in the hidden layer, hence creating a third new auto-association layer activation, and so on. The pattern obtained at the auto-association (input) layer after five re-injections is a RPP. Previous works (Ans and Rousset 1997, 2000; Ans et al. 2002, 2004) show that this number of re-injections was sufficient to correctly approximate the attractors and very little can be added by supplementary re-injections. Each component of a RPP can have any real value between 0 and 1. To display the selected RPPs, the real values were made discrete, using a 256-step linear scale that corresponds to the 256-greylevel scale generally used to display greylevel images (cf. Figure 4).

In order to generate the NIAs to be used with the opposition method in the simulation and in the behavioural experiments, it is crucial that the retained NIAs be more similar to the control than to the source items. Here are the selection rules that were used to select NIAs among the set of 4,325,000 RPPs that was initially generated:

*R*1- in terms of Euclidean distance, a selected NIA is closer to a control item than to any source item;

*R*2- the mean of the Euclidean distances between each source item and the centroïd of the selected base of NIAs (the mean NIA) is greater than the mean of the distances between each control item and the NIA centroïd.

Applying *R*1 and *R*2 leads to retention of 6.9% of the initially generated set of RPPs. Also, with the aim of reducing the number of NIAs to be used in the simulation and the second behavioural experiment while increasing their variety, the following additional rule was used:

*R*3- the distance (in term of the root mean squared error, RMS) between any two selected NIAs is greater than 0.15.

When the source list for generating the NIAs was List A, this procedure led to NIA_A, a 3000-NIA base. It is, however, important to ensure that any effect that would be found does not come from an item-in-list bias, that is, from serendipitous peculiarities of the items that were assigned to List A. To rule out this possibility, an NIA base counterbalancing was used: a 3000-NIA base NIA_B was generated by applying the same selection procedure to RPPs generated in a network trained this time with List B as source list (cf. Figure 2).

For each of these two selected NIA bases, it was first verified that the rule *R*2 related to the mean NIA held again. It was also verified that the following crucial property, expected from *R*1 and *R*2, was satisfied: the distance between any NIA and any control item was on average less than the distance between any NIA and any source item. In short, the NIAs are more similar to the control items at the exemplar level. So, even though the selected NIAs are truly very different from all the items (cf. Figures 2 and 3), they are yet more similar to the control than to the source items, both at the exemplar and at the centroïd level.

2. Simulation

The aim of this section is to check whether the memory of a distributed network, NET1, can be transported to another, initially untrained, neural network, NET2, only by NIAs, that is, by attractors that are not the source items (i.e. the items NET1 was trained on). The opposition method presented above was used to select the NIA training base. In case any memory of the original items could be evidenced using NIAs that are so different from them, two different result patterns are possible at test.

If the NIAs transport important pieces of information on NET1's 'internal landscape' to NET2, that is, if NIAs convey distributed information, we expect NET2 to perform better at test on the source than on the control items. On the contrary, if there is not such a thing as distributed information then owing to the constraints introduced because of the opposition method – which make NIAs more similar to the control items – NET2 is expected to perform best on the control items. The simulation was run to check which outcome occurs, and also to serve as a point of comparison for the behavioural results that are presented in the next section.

2.1. *Material and procedure*

Two different NET2 architectures were used in the simulation, to test for the generality and robustness of the results. The first one was identical to the one used to generate the RPPs, that is, a backpropagation auto-associator with a 169-unit auto-association layer and a 16-unit hidden layer (for further details see the previous section). The second architecture was identical except for the hidden layer, which contained 250 hidden units. We present these two architectures since they constitute two significant points in the continuum of the possible size of the hidden layer. Sixteen corresponds to the minimal number of hidden units that allows a good learning of the actual 53 patterns in NET1. On the other hand, 250 corresponds to the number of hidden units allowing to

obtain an RMS inferior to 0.01 on a set of test items in NET2 after only one epoch of learning of each NIA. The choice of a *single* training epoch was made for comparison sake with respect to the forthcoming behavioural experiments where each NIA base will be presented only once. NET2, initialised with random connection weights uniformly sampled between -0.5 and 0.5 was trained on an NIA base (either NIA_A or NIA_B, according to the NIA base counterbalancing) for a single learning epoch. It was then tested on both source and control list items. Twelve replications per NIA base were run, each with a network having different set of initial connection weights. For each replication, the connection weights were the same in the two conditions that resulted from the NIA list counterbalancing.

2.2. *Results*

The results when NET2 is identical to the one used to generate the NIAs (i.e. with a 16-unit hidden layer) are presented first. The average error – RMS between the activation produced at the NET2 auto-association layer and the tested item – was dramatically smaller for the source than for the control list items, $F(1, 11) = 3,030.576$, $MS_E = 0.00000547$, $p < 0.0001$. Hence, though drastically selected in order to be closer to the control items than to the source items, the NIAs generated in NET1 still give rise in NET2 to a higher familiarity (i.e. a lower RMS error) with the source items (mean RMS $= 0.2515$, SD $= 0.0083$) than with the control items (mean RMS $= 0.2887$, SD $= 0.0062$). The NIA base effect was significant, with a lower-average error on source and control items when NET2 was trained on NIA_A base than on NIA_B base, $F(1, 11) = 99.997$, $MS_E = 0.0000211$, $p < 0.0001$. The interaction between the factors *type of items* (source vs. control) and *NIA base* (NIA_A vs. NIA_B) was also significant, $F(1, 11) = 10.138$, $MS_E = 0.00000329$, $p < 0.01$, with a higher difference between source and control items when NIA_A base was used than when NIA_B base was used.

When NET2 comprised a 250-unit hidden layer, the pattern of results was qualitatively identical to the one found with the previous architecture. As illustrated in Figure 5, after training on an

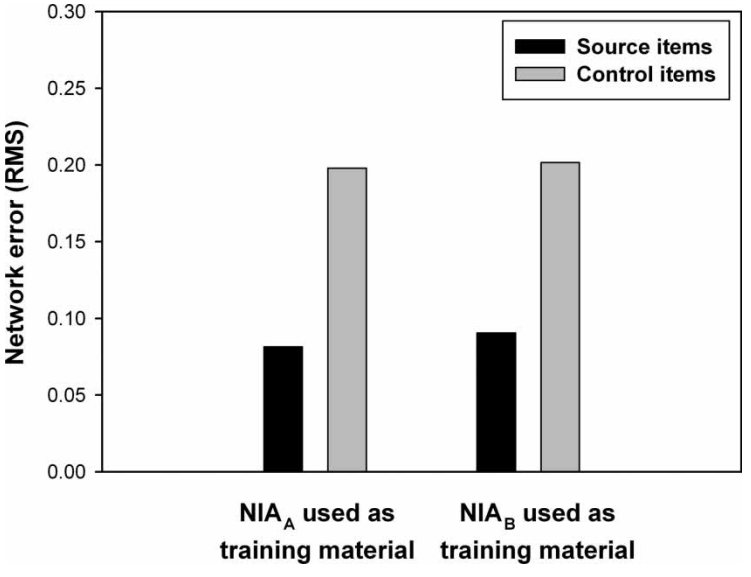

Figure 5. Network performance (RMS error) on source and control items after training on an NIA base (NIA_A or NIA_B, according to a source list counterbalancing). A lower RMS error signs a better performance.

NIA base, NET2 exhibited a higher familiarity with the source items (mean RMS = 0.0859, SD = 0.0051) than with the control items (mean RMS = 0.1996, SD = 0.0029), $F(1, 11) = 26,955.020$, $\mathrm{MS}_E = 0.00000576$, $p < 0.0001$. Also, a lower-average error on source and control items was found when NET2 was trained on $\mathrm{NIA_A}$ base than on $\mathrm{NIA_B}$ base – *NIA base* effect, $F(1, 11) = 94.767$, $\mathrm{MS}_E = 0.00000518$, $p < 0.0001$. The interaction between the factors *type of items* (source vs. control) and *NIA base* ($\mathrm{NIA_A}$ vs. $\mathrm{NIA_B}$) was also significant $F(1, 11) = 25.546$, $\mathrm{MS}_E = 0.0000036$, $p < 0.005$, with a higher difference between source and control items when $\mathrm{NIA_A}$ base was used.

NIAs can thus transfer specific information about the learned items. It is also to note that unsurprisingly they also transfer generic information about them: a test on random patterns that include the same proportion of black and white 'squares' as in control and source patterns produces a mean RMS value of 0.473 (SD = 0.025). This value, compared with that of 0.042 (SD = 0.010) and respectively of 0.297 (SD = 0.058) obtained, respectively, for source and control items clearly indicates that the network gained some knowledge of the common structure of the experimental pattern set induced by their construction algorithm: it does generalise to items it was not trained on when they come from the same 'family' (i.e. constructed by the same algorithm), but is at chance level for new items generated at random.

To sum up, the simulation yielded a consistent pattern of results, where the strongest effect was that of the type of items at test: the familiarity with the source items was consistently higher than with the control items. This means that NIAs not only are able to convey information from NET1 to NET2, but that this distributed information on the source items has a higher influence than the exemplar or centroïd information on the control items – with this latter having its origins in the selection rules. This result is a robust one, since it does not depend on the specific architecture of NET2, but on the nature of information conveyed by the NIAs. This bodes well for the behavioural experiments that follow.

3. Behavioural experiments

In the simulation above, NET2 trained on only NIAs exhibited at test more familiarity with the source than with the control items although the opposition method used to select the NIAs had as a consequence that the NIAs are more similar to the control than to the source items. So, the prediction issued from the simulation is that humans will, as NET2, exhibit at test more familiarity with the never seen source items than with the never seen control items. The general procedure for checking this prediction consists in presenting humans incidentally with selected NIAs generated in an auto-associator GDN that had previously been trained on a list of source items, then testing their familiarity with the never seen source items and the never seen control items.

In Experiment 1, the familiarity for the source items due to the NIAs will be measured by the number of false recognitions on the source items as compared with that on the control items, in a task measuring the recognition of some other, explicitly learned, target items. In a nutshell, the procedure is the following. First, participants perform a task during which they are incidentally exposed to selected NIAs then they are instructed to explicitly learn a list of target items. In the test phase, participants are to recognise the explicitly learned target items among distractors. Source and control items are used as distractors in this phase and should be rejected (i.e. not called 'learned'). Now, the more familiar an item, the more it will tend to elicit a 'learned' response, especially in an occurrence recognition task performed under time pressure like the one that will be used. Therefore, if humans are sensitive to distributed information conveyed by the NIAs, then the participants will be more familiar with the source items than with the control items and will thus tend to call the former 'learned' more often than the latter. To put it in other words, if humans

are sensitive to information conveyed by the NIAs, there will be more false recognitions on the source than on the control list items.

The reasoning at the root of the use of an occurrence recognition task in Experiment 1 is grounded in the dual-process model of recognition memory (Mandler 1980): Both familiarity and recollection of the original encoding episode contribute to recognition. More precisely, each old/new decision is based on a blend of an automatic process (called 'familiarity') and of a controlled, intentional process (called 'recollection'). Familiarity, being automatic, manifests itself more rapidly, so the blend is more familiarity-loaded for low-reaction times and gets more and more recollection-loaded as reaction times increase. This explains why under some conditions (e.g. limited response time) the recognition decision relies primarily (i.e. more heavily, yet not exclusively) on familiarity (Jacoby 1991; Ratcliff and McKoon 1995). It also explains why slower responses depend mostly (yet not exclusively) on recollection while familiarity plays a (much) lesser role. Here, because the recollection of the encoding episode will necessarily be spurious for never seen items, slower responses not only are less influenced by familiarity but depend on various response strategies that participants could bring into play in a recognition test. In the present experiment, given that the test items are items that were never seen before, recollection cannot contribute in an appropriate way to the answer to these items. So the conscious response strategies that go hand in hand with slower responses can just bring noise to the measure of familiarity. Thus, in accordance with Mandler's dual-process model of recognition memory, whose main reasoning was briefly exposed here, in order to favour contribution of familiarity in participants' responses only rapid responses will be analysed.

Experiment 2 aims to measure the familiarity component in isolation. After an incidental exposure to NIAs, the participants will perform a duration judgment task – under time pressure – both on the source and the control list items. Participants will be induced to believe that two slightly different presentation times are used and will have to classify items' display duration as *short* or *long*. Actually all items will have exactly the same duration. Participants' subjective impression that a given item's display duration 'is longer' is linked to an increased perceptual fluency (Jacoby 1983; Witherspoon and Allan 1985), whose real cause is familiarity with the item (Whittlesea, Jacoby and Girard 1990) – but that participants would attribute to different presentation times. Thus, if humans are sensitive to distributed information conveyed by NIAs, there will be a higher familiarity for the source items and thus more *long* responses on the source than on the control list items.

3.1. *Experiment 1*

Familiarity with the never seen items will be measured in this experiment through the false recognitions during an occurrence recognition task performed under time pressure. To do this, the following meaningful task was built up for the participants: after they were incidentally presented with NIAs, the participants learned a target item list and then had to recognise the target items presented among source and control items. This recognition task allowed testing whether previous exposure to NIAs resulted in more false recognitions on source than on control items.

3.1.1. *Method*

3.1.1.1 Participants. Forty-four students (mean age = 21.2 years, SD = 1.7) participated for course credit.

3.1.1.2 Stimuli. One hundred and five out of the 106 previously used matrices (cf. constructing the NIAs) were randomly divided into three lists: Target list (36 items), List 1 (35 items) and List 2

(35 items). Matrices were displayed as 260 × 260-pixel images centred on a black background. For each of the two groups resulting from the experimental design, NIAs were generated and selected according to $R1$ and $R2^2$ so as to build a 4500-NIA base. NIAs were displayed as 260 × 260-pixel matrices, like those depicted in Figure 3. All stimuli were displayed centred on a 17"screen (1024 × 768 pixels).

3.1.1.3 Design and procedure. Participants first performed an incidental study task: they were to detect a cross that appeared (9% of the trials) in a random location on a background made of NIAs: NIAs were displayed for 400 ms each, with no void in between. Prior to performing the task with the 4500-NIA base, participants underwent a warm-up phase where 500 of the 4500 NIAs were displayed.

The participants were then instructed to memorise the target item list (36 items) – in order to justify the later use of the recognition task to the participants. The target item list was presented six times with short pauses between presentations. For each presentation the target items were displayed in random order – each one for 500 ms, preceded by a black screen (1300 ms) and a fixation bar (200 ms).

Finally, the participants performed a recognition task where target, source, and control items appeared in random order. Items (1000 ms) were preceded by a black screen (2500 ms), a fixation bar (200 ms), and a black screen (200 ms). Participants were asked to press one of the two response keys 'as fast as possible, but without making haste errors' in order to make a *yes/no* recognition response within the 1000 ms of an item's display. This relatively low experimental cut-off of 1000 ms was chosen according to the following reasoning. On a theoretical ground (the dual-process model of recognition memory: Mandler 1980), the shorter the reaction time the larger the contribution of the familiarity process we are interested in – on the contrary, with larger reaction times the influence of the recollection process becomes more important and overwhelms that of the familiarity.

For counterbalancing sake, there were two experimental groups: The source list of Group NIA_{L1} was List 1 (and their control list was List 2), and the source list of Group NIA_{L2} was List 2 (and their control list was List 1).

3.1.2. *Results*

Positive recognitions (i.e. *yes* responses) made during the first 800 ms are considered. This analysis time limit was chosen in accordance with existing studies that have used a similar recognition task to assess familiarity (e.g. Jacoby 1991; Ratcliff and McKoon 1995) – and also because in this type of experimental paradigm responses given just around the experimental cut-off are necessarily noisy since they also reflect a strategic process (that related to the need to give a response, whatever it is, before the deadline). There was a clear difference between positive responses for the studied target items (9.52 responses, SD = 4.54) and positive responses for never seen items (4.41 responses, SD = 2.99), $F(1, 41) = 83.26, \mathrm{MS}_E = 767.046, p < 0.01$. This result indicates that 68% of the *yes* responses were correct recognitions (i.e. there were 32% false recognitions). Thus the participants performed adequately the recognition task, which implies that they had paid enough attention to the target items in the prior experimental phase and were able to memorise them, and later recognise them. This is an important corroboration because it ensures that the recognition task was meaningful for the participants. Having verified this, it makes sense to turn now to an analysis of the false recognitions.

Indeed, the results of interest concern the false recognitions: if, as predicted by the simulation results, the exposure to NIAs had induced a greater familiarity with the never seen source items than with the never seen control items, then more false recognitions would be observed for the

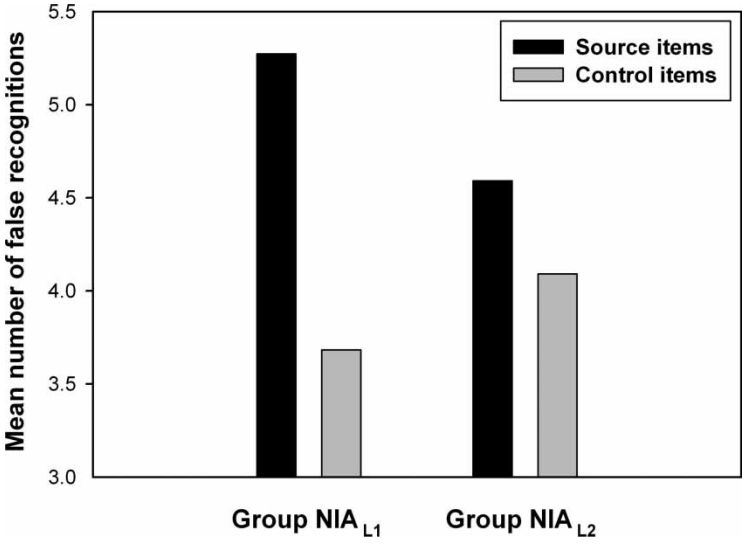

Figure 6. Experiment 1: false recognitions on source and control lists during the recognition task, after prior exposure to NIAs (Groups NIA_{L1} and NIA_{L2}, according to a source list counterbalancing). Note that, because of the design used, false recognitions are an indicator of familiarity with the never seen items.

source than for the control items. This is what was found: more positive recognitions (i.e. *yes* responses to never seen items, or false recognitions) were observed (cf. Figure 6) for the source (mean = 4.93, SD = 2.99) than for the control items (mean = 3.87, SD = 2.93), $F(1, 42) = 6.67$, $MS_E = 3.605$, $p = 0.0134$. No other effect was significant – group effect: $F(1, 42) = 0.029$; interaction: $F(1, 42) = 1.186$.

An equally important point to consider is whether this result is the consequence of familiarity or whether it arises from some artefactual reason. If the result is indeed grounded in participants' higher familiarity for the source items as compared with the control items, then a very similar pattern of results should occur when a lower cut-off is considered because familiarity influences predominantly fast responses (i.e. responses with short reaction times), especially, in experimental designs involving response deadlines (Mandler 1980; Jacoby 1991; Ratcliff and McKoon 1995; see also Yonelinas 2002) as the one used here. In the following, it is checked whether familiarity is indeed at play.

When other low cut-offs are considered, the result pattern remains the same. Restraining the analysis to those answers given within 700 ms, there are again more positive recognitions for the source (mean = 2.48, SD = 1.89) than for the control items (mean = 1.61, SD = 1.45), $F(1, 42) = 14.30$, $MS_E = 1.147$, $p < 0.01$. No other effect was significant – group effect: $F(1, 42) = 0.039$; interaction: $F(1, 42) = 0.357$. The same is true with an analysis cut-off of 600 ms: there were again more positive recognitions for the source (mean = 0.70, SD = 1.02) than for the control items (mean = 0.20, SD = 0.41), $F(1, 42) = 10.00$, $MS_E = 0.550$, $p < 0.01$. No other effect was significant – group effect: $F(1, 42) = 2.528$, $p = 0.119$; interaction: $F(1, 42) = 0.744$. In these cases too, the recognition task was meaningful for the participants, since out of the *yes* responses there were 67% and respectively 69% correct recognitions with a cut-off of 700 ms and respectively of 600 ms.

When considering the highest analysis cut-off possible given the response deadline in the recognition task, that is 1000 ms, the difference between the number of false recognitions for source (mean = 8.93, SD = 4.00) and for control items (mean = 7.77, SD = 3.91) is no longer significant, $F(1, 42) = 3.78$, $MS_E = 7.830$, $p = 0.062$ (no other effect was significant either – group

effect: $F(1, 42) = 0.993$; interaction: $F(1, 42) = 2.440$, $p = 0.126$). This reinforces the interpretation that due to the exposure to NIAs source items are more familiar than control items. Indeed, slower responses, given little before the deadline, are less influenced by familiarity and more influenced by strategic and conscious memory processes (that reduce the influence of the former), which may explain why with a cut-off of 1000 ms no reliable difference is found between source and control items.

As both source and control items have never been presented to the participants before the test phase, this pattern of results indicates that exposure to NIAs issued from of an artificial neural network induced greater familiarity with the never seen source items than with the never seen control items, though the selected NIAs are closer to the control items than to the source items – both at the exemplar and the centroïd level.

3.2. *Experiment 2*

Experiment 2 is strictly matched to the simulation with respect to the items and the NIA lists used. Contrary to Experiment 1, the use of a target list was avoided. The memory advantage for the never seen source items (over the never seen control items) is tested by comparing the perceptual fluency for source and for control list items.

3.2.1. *Method*

3.2.1.1 Participants. Seventy students (mean age = 20.5 years, SD = 1.6) participated for course credit. None of them participated in Experiment 1.

3.2.1.2 Stimuli. The original 106 simulation items were used, presented as 13×13 matrices (260×260-pixel images). As in Experiment 1, there was a 500-NIA incidental presentation warm-up base. Fifty-two NIAs were also used as a warm-up base for the perceptual fluency test. All apparatus details are those of Experiment 1.

3.2.1.3 Design and procedure. Participants performed the same incidental study task as in Experiment 1 – here, on a 3000-NIA base. They then engaged in a duration judgment task: they were to classify the display duration of images presented to them as *short* or *long*. In order to introduce the test to the participants progressively, a warm-up phase was designed. To prevent interference with test items, only NIAs were used during the warm-up phase. During the warm-up, the first 40 trials used two display durations (200 or 250 ms): After eight example trials, the participants received feedback on their responses to the remaining 32 trials. Twelve NIAs were then presented without feedback and with less different display durations (200 or 230 ms). After this warm-up, participants performed the experimental duration task, presented to them as 'the same test on a different type of stimuli'; unbeknown to the participants, the presentation time for the 106 items of interest (i.e. source and control items) was actually always of exactly 200 ms. The inter-stimuli interval was of 1300 ms. Participants were informed of the response deadline and had to answer within 1000 ms.

For counterbalancing sake, there were two experimental groups: the source list of Group A was List A (and their control list was List B), and the source list of Group B was List B.

3.2.2. *Results*

With a cut-off at 800 ms, there were (cf. Figure 7) more *long* responses to source (mean = 14.97, SD = 5.66) than to control items (mean = 13.84, SD = 6.31), $F(1, 68) = 4.517$, $\mathrm{MS}_E =$

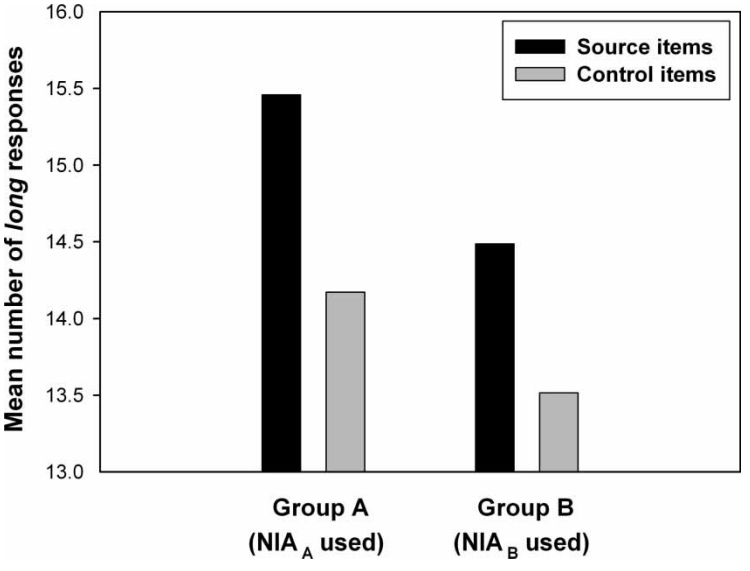

Figure 7. Experiment 2: effect of incidental exposure to NIAs on *long* responses (denoting familiarity) to source and control items in a duration judgment task (NIA_A was used for Group A and NIA_B for Group B, according to a source list counterbalancing). All stimuli correspond strictly to those used in Simulation.

9.868, $p < 0.05$ (no other effect was significant – group effect: $F(1, 68) = 0.370$; interaction: $F(1, 68) = 0.088$). Consistent with the idea that the above results are due to familiarity, when a lower cut-off of 700 ms was considered, the same pattern of results was observed: There were more *long* responses to source (mean = 13.56, SD = 6.34) than to control items (mean = 12.41, SD = 6.71), $F(1, 68) = 4.083$, $MS_E = 11.195$, $p < 0.05$; no other effect was significant – group effect: $F(1, 68) = 0.777$; interaction: $F(1, 68) = 0.092$. Because the task was more difficult than the recognition task used in Experiment 1, the participants needed more time to initiate a response, so there were too few responses to allow for a further analysis with a cut-off below 700 ms.

In line with the results from Experiment 1, strategic and conscious memory processes reduce the influence of familiarity on the slower responses: When the higher cut-off of 1000 ms is considered, the difference between the number of *long* responses to source (mean = 16.63, SD = 5.39) and to control items (mean = 15.60, SD = 5.95) does no longer reach significance, $F(1, 68) = 3.706$, $MS_E = 9.993$, $p = 0.057$. No other effect was significant either – group effect: $F(1, 68) = 0.502$; interaction: $F(1, 68) = 0.346$.

Taken together, these analyses show that the participants are more familiar with the never seen source items (than with the never seen control items), a stronger familiarity that is grounded in the prior exposure to NIAs.

4. Discussion

This paper deals with the question of whether humans have the ability to capture distributed information hold in a multi-layered network trained by a GDN procedure when presented with nonlinear attractors of the network (RPPs). The motivation for asking this question is that this ability is central to a model of human memory that supposes that long-term memory is related to a GDN (McClelland et al. 1995) and that memories have to be dynamically maintained by some memory self-refreshing mechanism as the reverberating process proposed by Ans, Rousset and collaborators (Ans and Rousset 1997; Ans and Rousset 2000; Ans et al. 2002, 2004; Ans

2004). Behavioural experiments directly answer this question: Humans can learn from distributed information.

Non items attractors (NIAs), that is RPPs selected so as to prevent them from being the exemplars used to train the network that generated the RPPs (under their initial form or as noisy, distorted versions) were used. In two behavioural experiments, it was found that humans are sensitive to information conveyed by the NIAs. In both experiments, even though the NIAs were more similar – both at the exemplar and the centroïd level – to the control items than to the items used to train the network that generated the RPPs (i.e. the source items), the participants presented exclusively with NIAs were shown to be more familiar with the source items than with the control items. What human memory properties are responsible for this *a priori* surprising result? We first discuss the contribution that the connectionist simulation makes to answer this question then we consider and discuss alternatives accounts.

The connectionist simulation was conducted using exactly the same training material (i.e. the same NIAs) as in Experiment 2 and used two different GDN auto-associator networks that differed markedly in their number of hidden layer units, and thus in their compression (i.e. the input to hidden units ratio). A first one had a compression ratio of more than 10, while the other's compression ratio was of 0.68. Whatever the compression ratio, the results of the simulation consistently yielded the same results as the behavioural experiments. Indeed, though trained only on RPPs selected to ensure that they are more similar to the control items than to the source items, the networks performed better at test on the never seen source items than on the never seen control items. Given the selection rules applied to select the NIAs among all the RPPs, this means that distributed information (on the source items) has a higher influence than exemplar or centroïd information (on the control items). The perfect match between the results of the simulation and those of human participants seem to point at the conclusion that the same is true for humans. Though participants may be sensitive to exemplar or centroïd information on the control items induced by the selection rules, distributed information on the source items held by the NIAs had an even greater influence on their performance.

Memory of the source items was evidenced after learning NIAs that were truly not distortions or noisy versions of them. The crucial feature of this study is that the selected NIAs are not only very different from the source items but also closer to the control items both at the exemplar and at the centroïd level. Here the selection rules were applied considering raw similarity in the physical space, but some cognitive models also refer to similarity within a psychological space (cf. Nosofsky 1992). Psychological space corresponds to a similarity metric that can be derived from categorisation or identification responses given by participants, and does not necessarily correspond to the physical space. In the present experiments, relations between physical and psychological spaces are thus to be considered. Following this psychological space argument, the main way to question the conclusions that we draw here on the basis of results obtained with the opposition method is to postulate that all the RPPs produced by the distributed neural network are actually closer to the source items within the human psychological space. The appreciable number of NIAs (about 7%) satisfying both distance selection rules ($R1$ and $R2$) that are at the root of the opposition method would then only reflect spurious discrepancies between physical and psychological spaces. While this possibility cannot be rejected on theoretical grounds, the simulation has also to be taken into account *per se*. For a neural network, we cannot refer to a pre-experimental psychological space and nevertheless it can generate a substantial number of NIAs that are, in its physical space, closer to control than to source items. This simple fact points out the limits of the psychological space argument. Indeed, in order to interpret differently the results presented here, one would have to suppose a very peculiar distortion between the physical space and the human psychological space. We consider that the facts that (a) source and control items are issued from a single and homogenous family of, (b) totally unknown patterns, (c) that was randomly divided in lists tend to make this particular distortion very unlikely. It seems more

likely that the results actually show that the psychological space of the participants was modified by a simple incidental exposure to thousands of NIAs closer to control items in a way that actually induces familiarity with the never seen source items.

It is necessary to acknowledge the fact that the results presented here are bound to auto-associations. This restriction has two reasons. The first one is of pragmatic nature: we could not imagine an experimental task where NIAs representing a hetero-association would be presented to the human participants and particularly could not come up with a meaningful test task to probe such hetero-associations. The second reason is that in neural network simulations, Ans and Rousset (2000) have convincingly shown that NIA-like RPPs issued from a mixed architecture containing both hetero and auto-associations can transfer not only memory concerning the auto-associative part but also the hetero-associative one, even though the RPPs were extremely far away from the initial patterns (the selected RPPs used were at least at a RMS of 0.5 from the initial patterns).

In themselves, the behavioural experiments presented in the paper do show that learning can occur in humans from NIAs. We acknowledge that because the visual modality was chosen to pass the NIAs to the human cognitive system and because of the particular experimental designs that were used, the present study does not allow one to decide whether learning affected the memory for earlier perceptual processes – e.g. perceptual representation systems (PRS: Schachter 1990; Tulving and Schachter 1990) – or occurred through changes in higher long-term memory systems.

Another topic that deserves discussion is that of the fundamentally different input that is given to the neural networks as opposed to the human cognitive system. More precisely, as an anonymous reviewer pointed out on a previous version of the paper, there is spatial information (i.e., connectivity between 'pixels') in the latter, while this topological information is absent in the input given to the neural networks. While we cannot but acknowledge this difference, it should also be pointed out that a perfect match between human perception and neural networks coding is very hard to achieve. Indeed, even if some invariant perceptual organisation principles are known, the importance of the topological situation and connectivity between 'pixels' for the humans is highly variable between individuals, as a function of the history of all the perceptual experiences of a given individual (e.g., Behrmann, Geng, and Baker 2005). Therefore, as there is no unique topological perception common to all humans (i.e. the weight that a participant would give to the fact that two 'pixels' are adjacent is different from the weight another participant would give to this information), it is not suitable to hard-code the spatial information available to the humans into the input given to the networks. What are the consequences of leaving aside this topological information? From a neural network point of view, the topological information being absent, the information the network would process is, on the one hand, that on the general deep structure common to all the training exemplars, and, on the other hand, that on the particular features of each training exemplar (McClelland and Rumelhart 1985). On top of that, humans process the topological information available, and their perception is influenced by their perceptual organisation scheme. Topological information and perceptual organisation in the participants cannot be turned off. For our purpose, however, these components are only introducing noise. Taking this into account, it is just more surprising that the effect of distributed information still affects humans in the way we hypothesised. To a certain degree, this may also contribute to the observation of much more clear-cut data in the simulation than in the behavioural experiments.

The present study, while it does provide original results, is just one among the possible studies. Nevertheless the present results indicate that humans are able to learn never seen source items from RPPs generated by a fully distributed artificial neural network trained on these items even though these attractors were more similar to the control items than to the source items. These results provide further insights on the conceptualisation of memory processes by evidencing that humans are sensitive to a specific type of information, distributed information, conveyed by nonlinear system attractors that do not amount to learned exemplars or their centroïd.

Acknowledgments

This work was supported by a research grant from the European Commission (HPRN-CT-1999-00065) and by the French government (CNRS UMR 5105). We thank Amanda Sharkey and two anonymous reviewers for their suggestions that helped us to improve the quality of the present paper, and Alan Chauvin, Robert French, Dwight Kravitz, James L. McClelland, David Plaut, and Gautam Vallabha for valuable discussions and suggestions on previous drafts.

Notes

1. While the first rule considers the exemplar level, this second one corresponds to a constraint at the prototype level. However, because here no set of items or of NIAs were created by distortion from a prototype, there is no such a thing as a prototype. Nevertheless, for any set of items a centroïd can be computed *a posteriori*. With the centroïd as a *post hoc* prototype, this second constraint considers the selection with respect to the centroïd.
2. Rule $R3$ was not applied to the NIAs used in Experiment 1. Leaving out $R3$ results in keeping NIAs that are more similar one to another, a situation that has as a consequence a higher-experimental noise. This situation is neutral with respect to the hypothesis at test.

References

Ans, B. (2004), 'Sequential Learning in Distributed Neural Networks Without Catastrophic Forgetting: A Single and Realistic Self-Refreshing Memory can do it', *Neural Information Processing-Letters and Reviews*, 4, 27–32.

Ans, B., and Rousset, S. (1997), 'Avoiding Catastrophic Forgetting by Coupling Two Reverberating Neural Networks' *Comptes Rendus de l'Académie des Sciences Paris, Life Sciences*, 320, 989–997.

Ans, B., and Rousset, S. (2000), 'Neural Networks with a Self-Refreshing Memory: Knowledge Transfer in Sequential Learning Tasks Without Catastrophic Forgetting', *Connection Science*, 12, 1–19.

Ans, B., Rousset, S., French, R.M., and Musca, S.C. (2002), 'Preventing Catastrophic Interference in Multiple-Sequence Learning Using Coupled Reverberating Elman Networks', in *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, eds. W.D. Gray and C.D. Schunn, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 71–76.

Ans, B., Rousset, S., French, R.M., and Musca, S.C. (2004), 'Self-Refreshing Memory in Artificial Neural Networks: Learning Temporal Sequences Without Catastrophic Forgetting', *Connection Science*, 16, 71–99.

Behrmann, M., Geng, J., and Baker, C. (2005), 'Acquisition of Long-Term Visual Representations: Psychological and Neural Mechanisms', in *Dynamic Cognitive Processes*, eds. N. Ohta, C.M. MacLeod, and B. Uttl, Tokyo: Springer-Verlag, pp. 11–35.

Blackmon, J., Byrd, D., Cummins, R., Poirier, P., and Roth, M. (2005), 'Atomistic Learning in Non-Modular Systems', *Philosophical Psychology*, 18, 313–325.

Carpenter, G.A., and Grossberg, S. (1988), 'The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network', *Computer*, 21, 77–88.

French, R.M. (1992), 'Semi-Distributed Representations and Catastrophic Forgetting in Connectionist Networks', *Connection Science*, 4, 365–377.

French, R.M. (1994), 'Dynamically Constraining Connectionist Networks to Produce Distributed, Orthogonal Representations to Reduce Catastrophic Interference', in *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Hillsdale NJ: Lawrence Erlbaum, pp. 335–340.

French, R.M. (1997), 'Pseudo-Recurrent Connectionist Networks: An Approach to the "Sensitivity-Stability" Dilemma', *Connection Science*, 9, 353–379.

French, R.M. (1999), 'Catastrophic Forgetting in Connectionist Networks', *Trends in Cognitive Sciences*, 3, 128–135.

French, R.M., Ans, B., and Rousset, S. (2001), 'Pseudopatterns and Dual-Network Memory Models: Advantages and Shortcomings', in *Connectionist Models of Learning, Development and Evolution*, eds. R.M. French and J. Sougné, London: Springer, pp. 1–10.

Grossberg, S. (1987), 'Competitive Learning: From Interactive Activation to Adaptive Resonance', *Cognitive Science*, 11, 23–63.

Hebb, D.O. (1949), '*The Organization of Behavior*', New York: Wiley.

Hetherington, P., and Seidenberg, M. (1989), 'Is There "Catastrophic Interference" in Connectionist Networks?', in *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, Hillsdale NJ: Lawrence Erlbaum, pp. 26–33.

Hinton, G.E. (1989), 'Connectionist Learning Procedures', *Artificial Intelligence*, 40, 185–234.

Jacoby, L.L. (1983), 'Perceptual Enhancement: Persistent Effects of an Experience', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 21–38.

Jacoby, L.L. (1991), 'A Process Dissociation Framework: Separating Automatic from Intentional Uses of Memory', *Journal of Memory and Language*, 30, 513–541.

Kortge, C.A. (1990), 'Episodic Memory in Connectionist Networks', in *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, Hillsdale NJ: Lawrence Erlbaum, pp. 764–771.

Lewandowsky, S. (1991), 'Gradual Unlearning and Catastrophic Interference: A Comparison of Distributed Architectures', in *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, eds. W.E. Hockley and S. Lewandowsky, Hillsdale NJ: Lawrence Erlbaum, pp. 445–476.

Lewandowsky, S. (1994), 'On the Relation Between Catastrophic Interference and Generalization in Connectionist Networks', *Journal of Biological Systems*, 2, 307–333.

Lewandowsky, S., and Li, S.C. (1995), 'Catastrophic Interference in Neural Networks: Causes, Solutions, and Data', in *New Perspectives on Interference and Inhibition in Cognition*, eds. F.N. Dempster and C. Brainerd, New York: Academic Press, pp. 329–361.

Mandler, G. (1980), 'Recognizing: The Judgment of Previous Occurrence', *Psychological Review*, 87, 252–271.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995), 'Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory', *Psychological Review*, 102, 419–457.

McClelland, J.L., and Rumelhart, D.E. (1985), 'Distributed Memory and the Representation of General and Specific Information', *Journal of Experimental Psychology: General*, 114, 159–197.

McCloskey, M., and Cohen, N.J. (1989), 'Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem', in *The Psychology of Learning and Motivation* (Vol. 24), ed. H.G. Bower, New York: Academic Press, pp. 109–165.

McRae, K., and Hetherington, P.A. (1993), 'Catastrophic Interference is Eliminated in Pre-trained Networks', in *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, pp. 723–728.

Murre, J.M.J. (1992), 'The Effects of Pattern Presentation on Interference in Back-propagation Networks', in *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, pp. 54–59.

Nosofsky, R.M. (1992), 'Similarity Scaling and Cognitive Process Models', *Annual Review of Psychology*, 43, 25–53.

Ratcliff, R. (1990), 'Connectionist Models of Recognition and Memory: Constraints Imposed by Learning and Forgetting Functions', *Psychological Review*, 97, 285–308.

Ratcliff, R., and McKoon, G. (1995), 'Bias in the Priming of Object Decisions', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 754–767.

Robins, A.V. (1995), 'Catastrophic Forgetting, Rehearsal and Pseudorehearsal', *Connection Science*, 7, 123–146.

Robins, A.V. (1996), 'Consolidation in Neural Networks and in the Sleeping Brain', *Connection Science*, 8, 259–275.

Robins, A.V., and McCallum, S. (1998), 'Catastrophic Forgetting and the Pseudorehearsal Solution in Hopfield-Type Networks', *Connection Science*, 10, 121–135.

Robins, A.V., and McCallum, S. (1999), 'The Consolidation of Learning During Sleep: Comparing the Pseudorehearsal and Unlearning Accounts', *Neural Networks*, 12, 1191–1206.

Schachter, D.L. (1990), 'Perceptual Representation Systems and Implicit Memory: Toward a Resolution of the Multiple Memory Systems Debate', in *Development and Neural Basis of Higher Cognitive Functions*, ed. A. Diamond, New York: New York Academy of Sciences, pp. 543–571.

Sharkey, N.E., and Sharkey, A.J.C. (1995), 'An Analysis of Catastrophic Interference', *Connection Science*, 7, 301–329.

Tulving, E., and Schachter, D.L. (1990), 'Priming and Human Memory Systems', *Science*, 247, 301–306.

Whittlesea, B.W., Jacoby, L.L., and Girard, K. (1990), 'Illusions of Immediate Memory: Evidence of an Attributional Basis for Feelings of Familiarity and Perceptual Quality', *Journal of Memory and Language*, 29, 716–732.

Witherspoon, D., and Allan, L.G. (1985), 'The Effect of a Prior Presentation on Temporal Judgments in a Perceptual Identification Task', *Memory and Cognition*, 13, 101–111.

Yonelinas, A.P. (2002), 'The Nature of Recollection and Familiarity: A Review of 30 Years of Research', *Journal of Memory and Language*, 46, 441–517.