scientific reports

OPEN

Comparing feedforward and recurrent neural network architectures with human behavior in artificial grammar learning

Andrea Alamia[1]✉, Victor Gauducheau[1], Dimitri Paisios[1,2] & Rufin VanRullen[1,3]

In recent years artificial neural networks achieved performance close to or better than humans in several domains: tasks that were previously human prerogatives, such as language processing, have witnessed remarkable improvements in state of the art models. One advantage of this technological boost is to facilitate comparison between different neural networks and human performance, in order to deepen our understanding of human cognition. Here, we investigate which neural network architecture (feedforward vs. recurrent) matches human behavior in artificial grammar learning, a crucial aspect of language acquisition. Prior experimental studies proved that artificial grammars can be learnt by human subjects after little exposure and often without explicit knowledge of the underlying rules. We tested four grammars with different complexity levels both in humans and in feedforward and recurrent networks. Our results show that both architectures can "learn" (via error back-propagation) the grammars after the same number of training sequences as humans do, but recurrent networks perform closer to humans than feedforward ones, irrespective of the grammar complexity level. Moreover, similar to visual processing, in which feedforward and recurrent architectures have been related to unconscious and conscious processes, the difference in performance between architectures over ten regular grammars shows that simpler and more explicit grammars are better learnt by recurrent architectures, supporting the hypothesis that explicit learning is best modeled by recurrent networks, whereas feedforward networks supposedly capture the dynamics involved in implicit learning.

In recent years the field of neural networks has undergone a substantial revolution boosted by deep learning approaches[1]. Different architectures have reached human-like performance in domains that were previously considered as sole prerogative of the human brain, such as perception[2] or language[3]. Part of this success originates from insights provided by cognitive sciences, in which brain-inspired solutions are implemented in functional models[4]. Conversely, it is possible to investigate the computational processes that take place in the human brain by comparing them with artificial functional models[5]. For this purpose, Artificial Grammar Learning (AGL) represents an ideal venue, given its well-established roots in both the cognitive and computer science literature. On the one hand, a formal definition of grammar complexity (i.e. Chomsky's hierarchy[6]) provides a theoretical framework to study grammar learning; on the other hand, previous studies in humans set a well-defined experimental framework to compare human behavior with the performance of different neural network architectures.

Formal language theory and Chomsky's hierarchy. Formal Language Theory (FLT) stemmed from studies grounded in mathematics and computability theory, realized in the first half of the previous century by logicians such as Emil Post or Alan Turing[7, 8]. In FLT, a *language* is defined as an infinite set of *strings*, which is in turn a sequence of symbols. Whether a string belongs or not to a language is determined by a *grammar*, i.e. a set of rules, and the distinction between grammatical and non-grammatical sequences is categorical (see Fig. 1A for examples). Noam Chomsky was the first to introduce a hierarchical classification of grammars composed of 4 nested levels, sorted on the basis of computational complexity[6,9]. The lowest level of the scale corresponds to *regular expressions*, which can be correctly classified by a memory-less finite-state automaton; the highest

[1]CerCo, CNRS, 31055 Toulouse, France. [2]Laboratoire Cognition, Langues, Langage, Ergonomie, CNRS, Université Toulouse, Toulouse, France. [3]ANITI, Université de Toulouse, 31055 Toulouse, France. ✉email: andrea.alamia@cnrs.fr
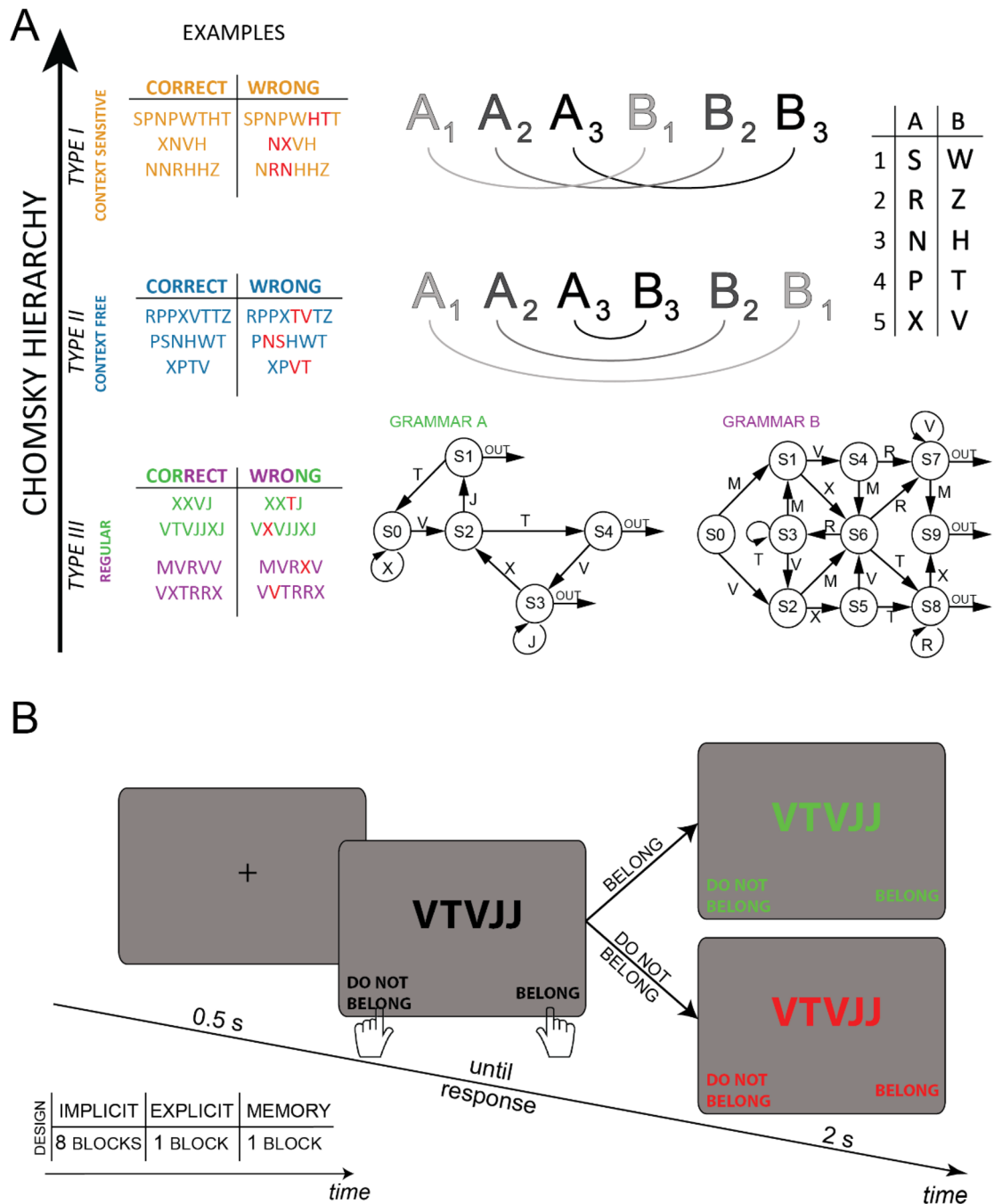
Figure 1. (**A**) A schematic representation of the 4 grammars employed in this study, arranged according to the Chomsky hierarchy. Type I or **Context Sensitive** grammar (CS, in orange), a translation was applied from the first to the corresponding second half of the string. For example, considering the first half as 'PPN' and referring to the same pairing as in the picture, the corresponding mirrored version would be 'HTT' in the CF, and 'TTH' in the CS grammar. The incorrect sequences were obtained by switching only two non-identical letters within one of the two halves (in red in the table). In type II or **Context Free** grammar (CF, in blue), the second half of the sequence mirrored the first half, applying the respective pairing to each letter (e.g. if A3 was the letter N, then the letter B3 was H). Note that a correct string in the context free grammar is considered as incorrect in the context sensitive, and vice versa. **Regular** grammars are defined by a scheme organized in directionally connected nodes (type III, the lowest in green and purple). For both grammars A and B, each correct sequence was generated starting at node S0 of each respective scheme, and then randomly choosing an arc departing from that node and recording its letter. This procedure was iterated until one of the arc labeled as 'out' was reached, thus terminating the string. Incorrect sequences had only one inexact item, being substituted by another letter from the proper vocabulary but violating grammatical rules (in red in the table). (**B**) Time-course of each trial. All the human experiments employed the same design. The sub-table reports the number of blocks for each session: implicit, explicit and memory. See "Experimental design" section for details.

level coincides with the *enumerable languages*, correctly classified by Turing machines. In between, *context free* (CF) and *context sensitive* (CS) grammars are respectively solved by push-down and linear-bounded automata[10]. However, this 4-levels ranking falls short when applied to natural languages (e.g. English). Consequently, one intermediate level has been added to the original hierarchy between CF and CS grammars, namely the *mildly context sensitive* grammar[11], which supposedly includes all natural languages[12]. However, experimental results have shown that participants can learn artificial grammars equally well irrespective of their level in the Chomsky hierarchy. This demonstrates that the hierarchy may reflect some form of computational complexity, but it does not reflect cognitive complexity[13]; possibly, different computational processes may be involved in human grammar learning.

Cognitive theories of artificial grammar learning.

Several studies have demonstrated that humans can perform above chance in artificial grammar learning paradigms. However, it is still debated what determines participants' behavior, and different theoretical accounts have been proposed[14]. The first theoretical description was provided by Reber[15], who suggested that participants learn the grammar' rules implicitly. Successive studies have questioned this interpretation, on the grounds of experimental evidence pointing at several potential confounds[16–18]. A revised account of the same hypothesis suggested that participants do not learn the full spectrum of grammatically correct transitions, but only a subset of it, i.e. microrules, such as bigrams or trigrams[19,20]. Similarly, the chunking hypothesis suggests that the more frequent chunks are the more salient, and consequently better learnt by participants as grammatically correct[21,22]. A further computational perspective posits that humans learn sequences as recurrent models, that is, by decoding relevant features from former adjacent items in the string[23–25]. Overall, contrary to human experiments in which subjects typically see a few dozen examples, all considered computational models have been trained with large datasets, and sometimes with significant overlap between training and test sets, making it difficult to draw any substantial comparisons with human cognition.

Implicit and explicit learning.

Another important aspect investigated in the experimental AGL literature regards the distinction between implicit and explicit learning. There are aspects of knowledge for which we have little conscious introspection, such as language[26]. Experimental evidence has suggested that AGL occurs implicitly, that is in an automatic and nonintentional way[27–29]. However, there are studies that demonstrated the implicit acquisition of knowledge in the framework of intentional learning, i.e. voluntarily modifying behavior in response to some feedback. An example of implicit intentional learning is the experiment by Berry and Broadbent[30], in which participants learnt to optimize the outcome of a complex dynamical system, without knowing explicitly the underlying rules and dynamics of the system[30,31]. Despite skepticism about the very existence of implicit learning ([16,32] but see in opposition[33]), the existence of distinct implicit and explicit systems in the human brain has been postulated, the former being automatic, the latter being flexible and directed to hypothesis testing[34,35]. In addition, previous AGL experiments demonstrated that more complex grammars are more likely to be processed implicitly than simpler grammars[36–38]. This result is in accordance with the assumption that working memory has restrictions on the number of items (i.e. hypotheses) explicitly accessible at any given time[39,40], thus limiting abilities to process complicated rules (e.g. based on long-distance dependencies). On the contrary, implicit processes possibly rely on different mechanisms, capable of dealing with larger amount of information at the cost of reduced flexibility[41–43].

Purpose of the study.

In this study we tested 4 grammars spanning over 3 Chomsky's hierarchy levels. Both human participants and artificial neural networks were trained and tested on datasets generated from those grammars. Importantly, we aimed to use comparable amounts of training for humans and artificial neural networks, which were composed of fully-connected layers, whose parameters were trained via backpropagation (through-time in case of recurrent networks). Our purpose was to investigate which architecture—feedforward vs. recurrent networks—better captures human behavior as a function of grammar complexity. Moreover, as AGL is an established framework to contrast implicit and explicit learning[44,45], we aimed at testing whether these modes could be related respectively to feedforward and recurrent architectures, similarly to findings in visual perception[46,47].

Materials and methods

Artificial grammar datasets.

We performed 4 experiments with different artificial grammars (Fig. 1A), each composed of the same amount of correct and incorrect sequences. According to the Chomsky hierarchy introduced above, two grammars were *regular*, referred in the following as grammar A and grammar B, one was *context free* and one was *context sensitive*. In the **regular** grammars, the sequences' length ranged from 2 (grammar A) or 3 (grammar B) to an arbitrary maximum length of 12 items. Longer sequences were discarded. Grammar A's vocabulary was composed of 4 letters, while grammar B counted 5 letters. The dataset from **context free** and **context sensitive** grammars (respectively type II and type I of the Chomsky hierarchy) were composed following an approach based on symmetry[10,13]. In both cases the vocabulary was composed of 10 letters combined in 5 pairs, and the strings' length was either 4, 6 or 8. Figure 1A shows in details how sequences were generated in each grammar.

Humans.

Participants. Overall, 56 participants (31 female, age = 25.4 ± 4.7) took part in 4 experiments using 4 different artificial grammars (n = 15 for Grammar A, B and Context Free; n = 11 for the Context Sensitive grammar). All participants gave written informed consent before the experiment, in accordance with the Declaration of Helsinki and received monetary compensation. This study was carried out in accordance with

the guidelines for research at the "Centre de Recherche Cerveau et Cognition" and the protocol was approved by the committee "Comité de protection des Personnes Sud Méditerranée 1" (ethics approval number N° 2016-A01937-44).

Experimental design. The same experimental design was applied for each grammar. Each trial started with a fixation cross lasting 500 ms, followed by a string of letters displayed in the center of the screen (Fig. 1B). Participants were informed that there were two groups of respectively correct and wrong sequences. They were asked to classify each sequence by pressing one of two key-buttons (respectively with the right and left index). Participants were not explicitly instructed about the existence or the nature of the rules generating the sequences. No time constraints were imposed to provide an answer. As soon as the response was given, visual feedback was provided: the string turned green when correctly classified, red otherwise. The visual feedback lasted for 2 s before starting the next trial. At each trial we recorded accuracy and reaction times.

Each participant performed one session lasting approximately 1-h and composed of 10 blocks. During the first 8 blocks, labeled as *implicit* in the sub-table in Fig. 1B, participants were not explicitly informed about the existence of the rules generating the sequences. Each block of the *implicit* part counted 60 trials, for 480 trials in total. A questionnaire was provided between the 8th and the 9th block to assess participants' explicit knowledge of the rules. The questionnaire was different for each grammar, asking specific questions about the rules (see appendix C). In grammar A and B participants responded to 7 multiple-choice questions, whereas in CF and CS grammars participants were asked to point to the wrong letter in a series of 7 novel sequences. In all grammars, participants were asked to report their confidence level from 0 to 100 after each answer. Following the questionnaires participants were asked to report (part of) the rules at the best of their knowledge. The last 2 blocks (labeled respectively as *explicit* and *memory* in Fig. 1B) were identical to the previous ones, but served as control conditions. In the 9th block, composed of only 20 trials, participants were provided with a printed scheme explaining exactly the rules of the grammar (the same as reported in Fig. 1A), and were instructed about the generation of correct sequences. During this block, they were allowed to consult the scheme before providing each answer. In the 10th and last block, participants were asked to perform the same task for additional 20 trials but no longer had access to the grammar scheme, thus supposedly relying on their memory of the rules.

Artificial neural networks.

Experimental design. The neural network design was composed of two parts: a first parameter search, and a subsequent comparison with human behavior. Both were implemented using the Keras library[48], back-ended in Tensorflow[49]. Altogether, we trained feedforward and recurrent architectures, each composed of a series of fully connected layers. All networks were trained to classify the sequences as correct or wrong, employing the same dataset (i.e. 4 grammars) and the same amount of trials as in the human experiments.

Regarding the parameter search, we aimed at determining the parameters whereby each architecture scored closest to human performance. We tested a range of networks, varying the number of layers and the learning rate, defining a 2-dimensional space (see Fig. 3). The training set was composed of 500 sequences (roughly similar to humans, who viewed 480 training examples), whereas the validation and testing set were composed respectively of 100 and 200 sequences. All the layers of a given network counted the same number of neurons except the output layer, which had only one neuron. The number of neurons was chosen such that all networks within the same 2-dimensional space had (roughly) the same number of free parameters. We explored 4 possible spaces with different numbers of parameters: respectively 1400, 7900, 31,000 and 122,000. One axis of the space referred to the number of layers, counting 6 levels (i.e. 2, 3, 4, 5, 7 and 10), whereas the other axis represented the learning rate, which counted 20 levels. Different values were used for the two families of architectures (see below). Each parameters space counted $6 \times 20 = 120$ networks, each one trained 20 times with random weights initialization. At first, we determined which networks provided the closest-to-human performance. For each grammar, we averaged between subjects the performance on the last block, and we subtracted this value from the mean performance of each network computed over 200-sequences (test set) after a learning of 500 examples (training set). For each parameters-space and grammar, we selected the network with the smallest absolute difference as the one closest to human behavior. Note that the selected network is not necessarily the one with the highest performance on the test-set (see Fig. A1 in appendix A). Once we determined the closest-to-human networks, we obtained their respective learning curves by varying the training set size progressively from 100 to 500 sequences, with a stride of 100. As in the parameter search, we averaged the results over 20 random weights initialization, using respectively 100 and 200 sequences for the validation and test set.

Feedforward architectures. Feedforward neural networks were composed of fully connected dense layers. The input layer counted 12xK neurons, representing the one-hot encoding of the 12-letters longest possible string (K represents the total number of letters, equal to 4, 5 and 10 for grammar A, B and CFG/CSG respectively). We employed zero-right padding when shorter sequences were fed to the network. All activation functions were defined as rectified linear units, i.e. 'ReLu', except the output neuron, which was implemented with a sigmoid function. The loss function was defined as 'binary cross-entropy', and optimized by means of stochastic gradient descent[50] with Nesterov momentum[51] set to 0.9, and decay equals to 1e−06. In all grammars, both in the parameter search and in the learning curve estimation, we considered 1 epoch only (500 trials), with batch size of 15.

Recurrent architectures. Recurrent neural networks were composed of fully recurrent connected layers, in which each neuron was connected to itself and all other neurons in its layer. Starting from each sequence's first letter, at each time step the following letter was provided to the network as a one-hot encoded vector. The input layer thus counted as many neurons as letters in the grammars alphabet. A sigmoid activation function armed

the output neuron, whose activation determined the classification decision after the last letter of the string was fed to the network. Learning occurred via back-propagation through time[52,53]. As in the feedforward architecture, we considered 'binary cross-entropy' as loss function, optimized with the keras function rms-prop[54]. We set rho to 0.9, epsilon to 1e−8 and decay to 0. As for the feedforward networks, we employed only 1 epoch (500 samples) and batch size of 15 in both the parameter search and the learning dynamic part.

Data analysis. In the human experiments, we recorded accuracy and reaction times at each trial. Both measures were averaged over blocks and tested by means of Bayesian ANOVA[55,56]. Each analysis provides a Bayes Factor (BF) for each independent factor (e.g. BLOCK). As widely accepted[57], we considered BF above 3 as substantial evidence in favor of the alternative hypothesis, over 20 as strong evidence, and BFs beyond 100 correspond to very strong evidence. Conversely, a BF below 0.3 suggests lack of effects[58,59]. In all our analyses we reported BFs and estimate's error.

Results

Human results. *Accuracy and reaction times.* For each grammar we assessed whether participants learned the rules by testing accuracy and reaction times (RT) by means of a Bayesian ANOVA, considering BLOCK as independent factor (categorical, from 1 to 8) and SUBJECT as a random factor. At every level of the Chomsky's hierarchy, participants learned the rules above chance (all BF > >100 for the BLOCK factor, for each grammar), as shown in Fig. 2A. Conversely, reaction times did not show any significant effect at any level of the Chomsky hierarchy, remaining between 2 and 3 s during the whole experiment. We found very similar results in both accuracy and RT in analyzing the sub-set of participants who failed in reporting the rules during the questionnaire in CF and CS grammars (respectively N = 8 and N = 12, in darker colors in the figure). Regarding grammars A and B, all the participants failed in reporting any subset of the rules.

Furthermore, we performed a Bayesian ANOVA to contrast the *explicit* and *memory* blocks with the 8th block of the *implicit* session (categorical factor CONDITION, 3 levels), considering only the implicit sub-set of subjects in CFG and CSG, and all the participants in grammars A and B. This analysis revealed a significant difference between conditions at every level of the Chomsky's hierarchy (all BF > >10) with larger differences between *implicit* and *explicit* blocks. Moreover, the accuracy was similar between the memory and the explicit tasks in all level of the Chomsky hierarchy (all BF < <3) except grammar B (BF = 7272, error = 9e−9%), suggesting that once participants knew the rules they could recall them shortly afterward (full results table in appendix B, Table B.1). All in all, the difference between *implicit* and *explicit* blocks suggests that participants were not fully aware of the rules during the *implicit* blocks, as their behavior changed when they could inspect the grammars' scheme. This conclusion is further corroborated by the subsequent questionnaires' analysis.

Questionnaire. Participants filled a questionnaire after the 8th block of the experiment (end of the *implicit* part). Regarding the *regular* grammars A and B, each question conceded 5 answers, having more than one possibly correct (see appendix C1). For each question we computed the Sensitivity and the Specificity (see supplementary material, appendix B), defined as the proportion of positive (Sensitivity) or negative (Specificity) responses correctly identified as such[60]. Interestingly, we observed a significant difference between grammar A and B in the Sensitivity index (BF > 100, error = 7.69e−7), revealing a more explicit knowledge in grammar A than B (Fig. 2B). No difference was observed in the Specificity index (BF = 0.647, error = 0.003). Regarding CF and CS grammars, participants were first presented with a new series of 7 incorrect strings (i.e. only one letter violated the grammar), and they were asked to indicate the one wrong letter in each sequence. Next, they were asked to report the rules or pattern they followed to perform the task (see appendix C.2). Remarkably, we observed a very clear pattern of result in these tasks: participants who correctly reported the rules' scheme as represented in Fig. 1A (3 participants for both grammar CFG and CSG) were also able to detect the wrong letters in the 7 incorrect sequences provided in the questionnaire. Contrarily, participants who reported no rules or wrong ones, were unable to detect the wrong letters in the incorrect sequences (they provided either wrong or no answers). Even though the distinction between implicit and explicit learning may not be categorical[61], the clear-cut distinction we observed in participants' questionnaire responses persuaded us to analyze separately the subset of participants who failed in reporting the rules at the end of the experiment (labeled 'implicit' in Fig. 2, as opposed to 'all', in which we considered all participants irrespective of their responses in the questionnaire). Importantly, we included in the implicit group only those participants that failed both the letter-identification task and failed in reporting the rules, thus making our criterion as conservative as possible. Lastly, in all grammars we compared the averaged confidence report provided after each question (Fig. 2B). Overall, we did not observe any significant difference between the grammars (BF = 0.216, error = 4.6e−4; very similar results were obtained considering the subgroup of "implicit" participants—see above—in grammars type II and I).

Artificial neural networks results. Regarding the neural networks analysis, we first performed a parameter search in order to identify the networks whose performance was closest to the human one. We tested 120 networks for each parameter space (each one defined by 2 axes: learning rate and number of layers—see methods "Experimental design" section). Each network was trained with a similar number of trials as in the human experiment (i.e. 500 compared to 480 in humans), and then tested over 200 sequences (with frozen parameters, i.e. no training). Figure 3A shows the absolute difference between human performance in the last block (i.e. trials from 420 to 480) and the networks performance after a training period of 500 sequences. Only the results regarding the 31,000 parameters space are shown, but similar results were obtained considering all other spaces (see in appendix A figure A.1). Notice that the networks closest to human performance (i.e. whose absolute difference is the smallest) are not necessarily the one performing the best. Specifically, regarding the feedforward networks,
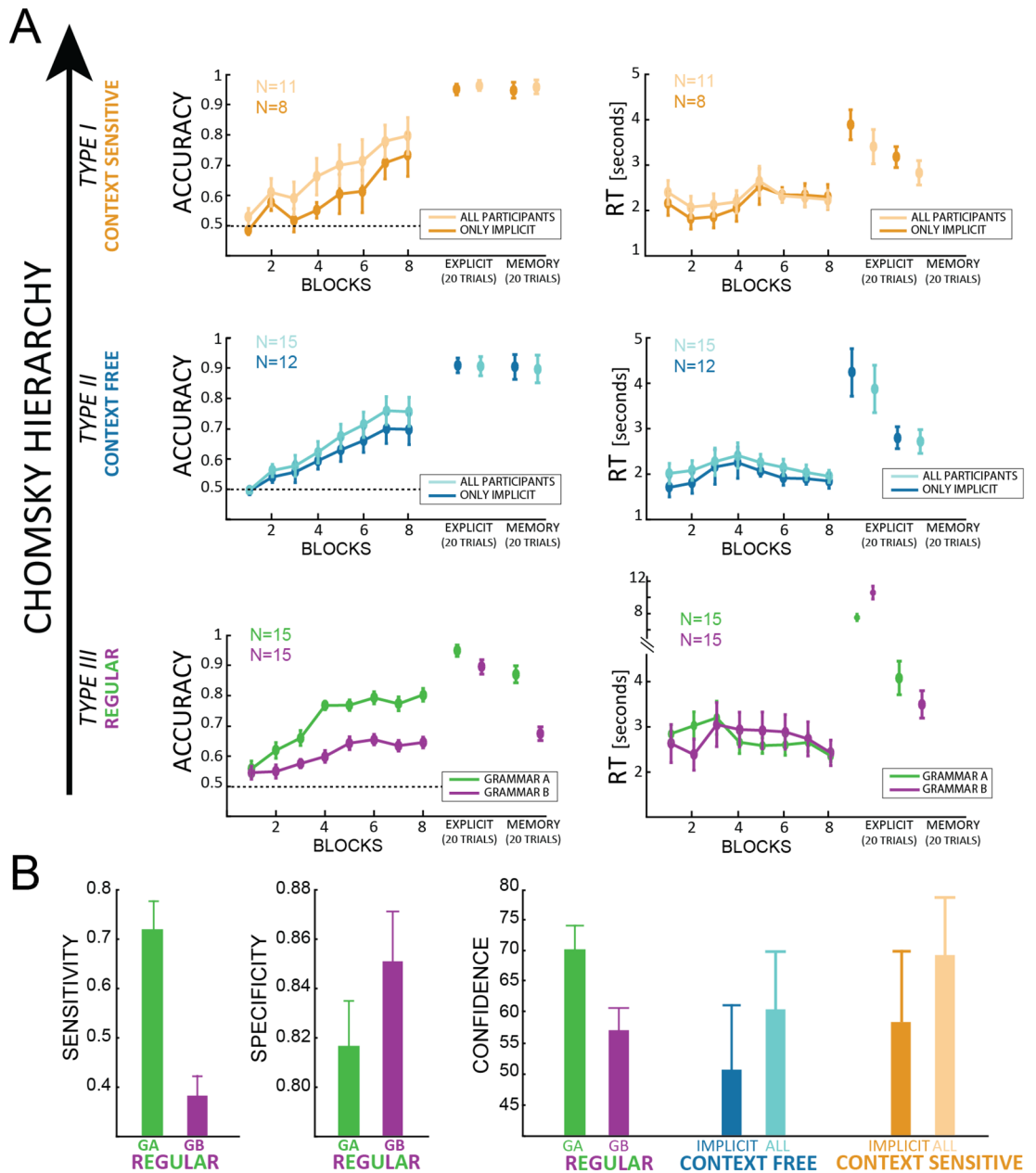
Figure 2. (**A**) For each level of the Chomsky hierarchy, accuracy (left column) and reaction times (right column) are plotted as a function of the implicit BLOCKS, followed by the EXPLICIT and MEMORY blocks. In type I and type II (first 2 rows), darker colors represent the sub-set of "implicit" participants (i.e. those who failed to report the rules during the questionnaire, see paragraph "Questionnaire" for more details), lighter colors all the participants. In type III, green and purple represent respectively grammar A and B. All values are expressed as average over participants ± standard errors (N values reported in each subplot). (**B**) Sensitivity and specificity indexes for explicit report of the grammar rules are expressed as average ± standard errors for grammar A (green) and B (purple). The right subplot shows the average confidence values expressed by all participants (darker colors in CFG and CSG represent the implicit sub-set of participants).

we observed that the best results were obtained with the lowest number of layers (corresponding to the highest number of neurons per layer), at each level of the Chomsky's hierarchy. Since the feedforward networks under-performed when compared to human performance, the ones with the best performance were also the closest to human's behavior. Concerning the recurrent networks, we observed that the higher accuracies corresponded to lowest learning rates, and that in regular grammars, the closest-to-human network did not correspond to the one with the best performance (the difference in performance between the best and closest-to-human networks are 0.02 and 0.08 for grammars A and B respectively). Additionally, we averaged the results from all the levels of the Chomsky hierarchy to determine which neural network was the closest to human behavior irrespective of the rules' complexity level (Fig. 3C,D—see appendix A figure A.2 for performance over trials). The results confirm that feedforward architectures perform best with the lowest number of layers, whereas recurrent architectures
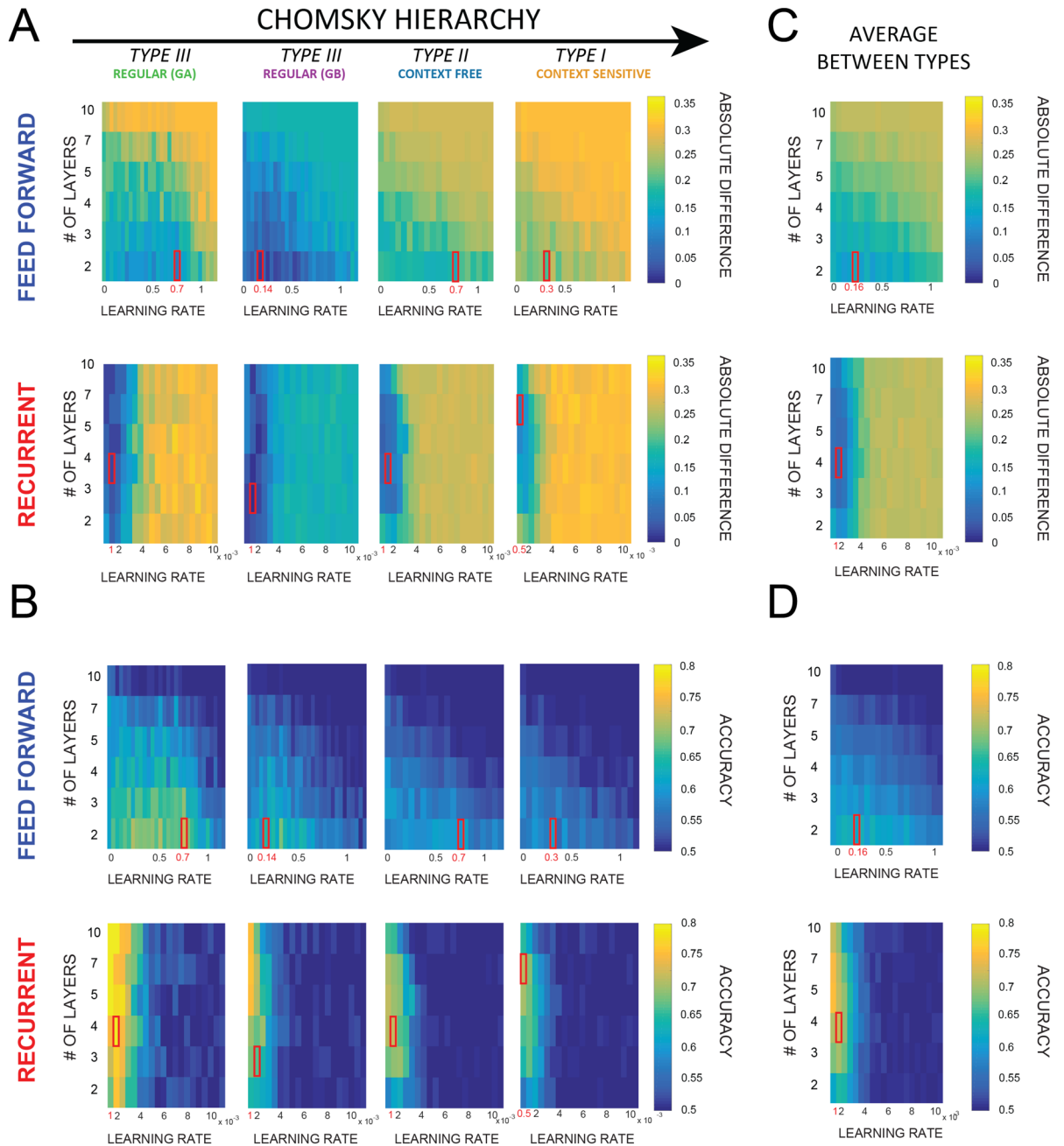
Figure 3. Results of the parameter search for the third space (i.e. 31,000 parameters). Each 2D space is defined by two axes: learning rate (LR) and number of layers (NL). Each point in the space represents the averaged performance of the network with specific LR and NL. The red square highlights the network closest to human performance for each plane, as the ones with the smallest difference (between NN and human performance) at the end of training, as shown in (**A**). In all subplots, upper rows are feedforward networks, lower rows are recurrent ones. (**A,B**) Each column represents one grammar, ordered according to the Chomsky's hierarchy. Subplot A shows the difference between NN and human performance, i.e. the distance to human performance at the end of training (see "Experimental design" section for details), subplot B the accuracy of each network. Similar results are observed between grammars: in FF networks 2 layers are sufficient to obtain the best results, whereas recurrent networks perform best with the lowest LR. (**C,D**) Same as (**A,B**) but considering the average across all grammars.

achieve the best accuracy with lower learning rate. Finally, we obtained the learning curves for the 8 best-performing models (4 grammars * 2 network types) in order to compare them with the human ones.

Human–networks comparison. To investigate whether not only the final accuracy, but also the temporal dynamics of training are comparable between humans and artificial neural networks, each network selected based on its final performance after 500 training trials (Fig. 3) was re-trained with 5 different training-set sizes,

ranging from 100 to 500 trials, and each time its performance was tested over 200 novel strings. The results are shown in Fig. 4 and represent the average ± standard error over 20 repetitions. For the human learning curves, we averaged a 40-trials long window centered over the trials corresponding to the same size as the training set. The first/last bins were computed averaging the first/last 40 trials of the first/last block. We performed a Bayesian ANOVA considering as categorical factors TRIAL (from 1 to 5, corresponding to the size of the training set), AGENT (three values representing humans, feedforward—FF—and recurrent—RR—architectures) and GRAMMAR (4 levels corresponding to the 4 grammars we tested). The results suggested a robust effect of each factor (all BF > > 3e+15, error < 0.01%), and a very strong interaction between the factors AGENT and GRAMMAR (BF = 2.3e+14, error < 0.01%). In order to disentangle this significant interaction, we tested a post-hoc analysis for each grammar separately, considering as factors only TRIAL and AGENT. Interestingly, in the highest level of the Chomsky hierarchy (type I and type II) we observed a significant difference between humans and FF (all BF > 1000, error < 1e−6) and between RR and FF (all BF > 1000, error < 1e−6)), but not between humans and RR (all BF < 1, error < 0.001). In type III grammars we reported two different patterns of results: in grammar A we observed results similar to the previous ones, having a significant difference between all the agents (human-FF and RR-FF: BF > 1000, error < 1e−6; human-RR, BF = 10.53, error = 2.5e−7). However, such a difference did not emerge in grammar B, in which we observed no difference between AGENTs (BF = 0.79, error = 0.02%). In summary, as shown in Fig. 4C, the pattern of results suggests that recurrent architectures are closer to human behavior at every level of the Chomsky hierarchy, with the exception of grammar B, for which recurrent and feedforward models cannot be distinguished, as further discussed in "Difference between feedforward and recurrent networks in regular grammars" section. This result was further corroborated by a ROC analysis, which confirmed that recurrent networks and human performance share a similar pattern of results, whereas feedforward networks, despite performing above chance, obtain significantly worse results (except on Grammar B; see supplementary figure A4).

Moreover, we also compared models and humans performance as a function of sequence length (Fig. 4B,D). For each grammar, we tested a Bayesian ANOVA having as factors sequence LENGTH, and AGENT (human, FF or RR). Interestingly, we found a strong effect of the factor LENGTH for each grammar (all BF > > 1000, error < 0.4%), revealing that human participants and both model architectures performed better with shorter sequences (table B.2 in appendix B for full results). Moreover, in all except grammar B we found a very significant difference between AGENTs (grammar B, BF = 0.38, error = 0.022%; all other grammars BF > 1000, error < 0.9%). Post-hoc analyses confirmed our previous results, revealing that recurrent architectures matched more closely human data than the feedforward ones. Lastly, it is noteworthy that vanishing gradients may be responsible for the decreasing performance with longer sequences, a problem that different implementations of recurrent and feedforward networks could prevent (i.e. Gated Recurrent Units or Long-Short Term Memory recurrent networks, but also residual or convolutional networks). However, an exploratory analysis using LSTM on the current dataset (maximum 500 examples, trained over 1 epoch) using the same parameters as the recurrent networks suggested that in our study (i.e. on these grammars and using these parameters) potentially more training examples would be necessary to properly train a gated version of our recurrent networks than the number of examples used by human subjects (see supplementary figure A5). On the other hand, a complete search in the parameters space (i.e. learning rate, number of layers/neurons) would may still provide a gated recurrent network (e.g. LSTM) with the ability to perform the task with the same limited number of training examples. We defer such an investigation to future work, and keep the focus of the present study on the plain comparison between the two main classes of architectures (i.e. feedforward and recurrent networks), rather than on each specific implementation.

Difference between feedforward and recurrent networks in regular grammars.

Previous results suggested that recurrent architectures perform better in modelling human learning than feedforward ones at every level of the Chomsky hierarchy. However, one of the regular grammars we tested (i.e. grammar B) violates this conclusion, revealing no significant differences between the two architectures and the human behavior. In order to shed some light on this result, we collected 8 additional artificial regular grammars from a recent review[62] and tested both our recurrent and feedforward architectures on each grammar. Based on our data, in which a simpler grammar (GA) leads to a larger difference between FF and RR networks than a more complicated one (GB), we hypothesized that RR networks would perform better (i.e., closer to humans) than FF networks in simpler grammars. Consequently, we defined 5 simple metrics to characterize the complexity of each grammar: number of letters in the vocabulary; number of states; number of transitions (or rules); number of bigrams and shortest length of a sequence (or minimum length).

Overall, the recurrent network always performed better than the feedforward one for all 10 grammars (Fig. 5A, upper panel). However, the difference between the two architectures was not constant across grammars. As shown in Fig. 5B, such difference correlated significantly (and negatively) with each of the complexity metrics (all BF > 5, error < 0.01%), suggesting that the difference between the two architectures is inversely correlated with the complexity of the grammars (see in appendices figure A.3 and table B.3 for the correlation matrix). In other words, over ten different regular grammars, the recurrent networks systematically outperform the feedforward ones (Fig. 5A), but less so in more complicated grammars; this was verified by a negative correlation between the recurrent-feedforward difference in accuracy and different complexity indexes (Fig. 5B). To confirm this conclusion, we created two novel original grammars (Fig. 5C), one at each extreme of the complexity metrics we defined above, and we trained feedforward and recurrent models on these new grammars. As predicted, the difference between FF and RR was large (0.11%) in the simplest and small (0.01%) in the more difficult grammars (numbered respectively 11 and 0 in Fig. 5A). Taking together the results of these simulations and the results of the questionnaire in the regular grammars (Fig. 2B, left panels), which suggest that simpler grammars are learnt
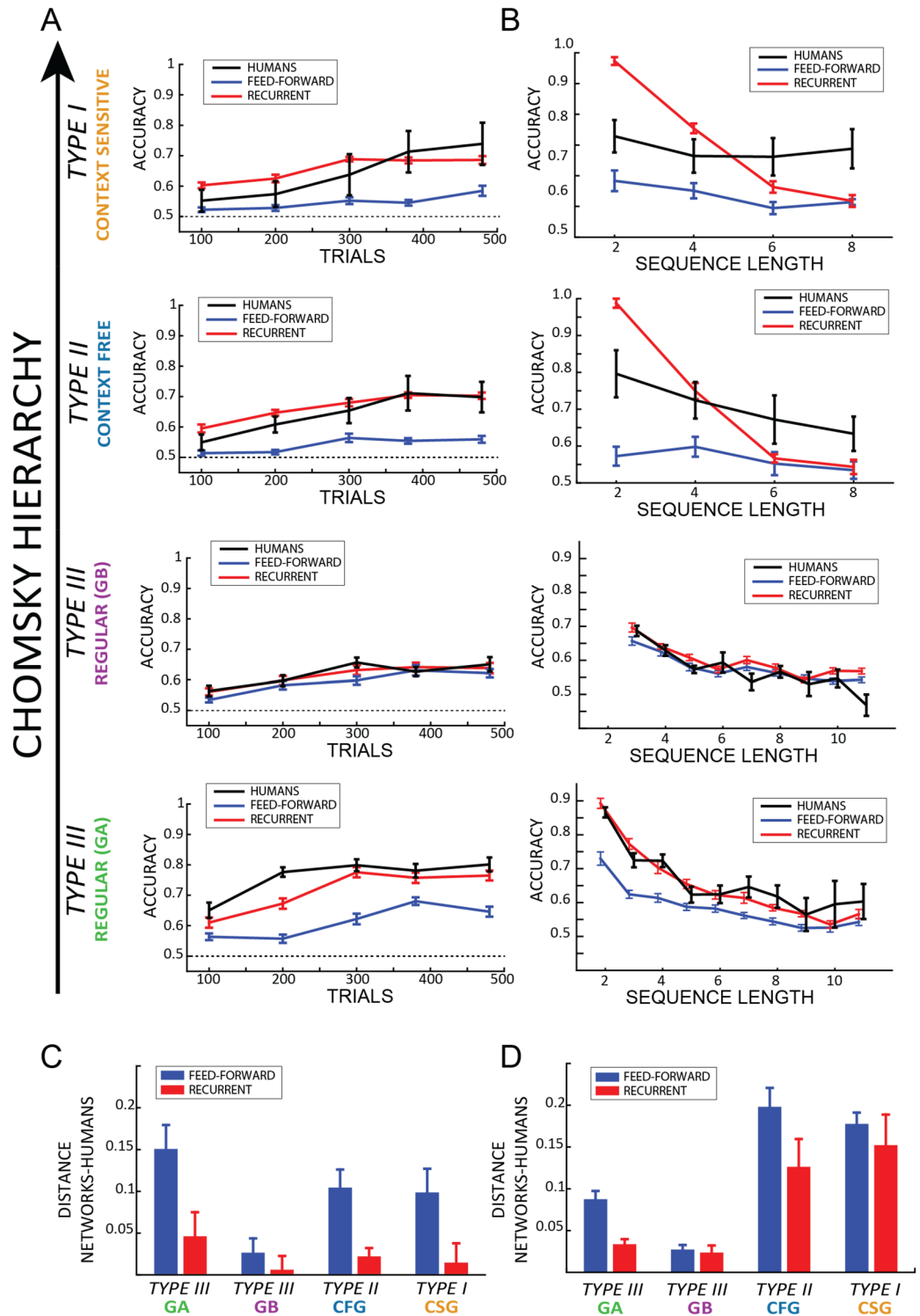
Figure 4. (**A**,**B**) Results over trials (**A**) and string lengths (**B**) for humans (in black) feedforward (in blue) and recurrent (in red) networks. For humans in (**A**) each bin is an average over 40 trials (20 trials before and after respectively, except the last bin which includes the last 40 trials of the experiment). Each row represents a grammar, ordered according to the Chomsky's hierarchy. (**C**,**D**) The plots show the distance between humans' performance and FF (in blue) and RR (in red). Each distance represents the area measured between the human and the network curves. Except for grammar B, RR networks are significantly closer to human performance, both as a function of training time and sequence length.
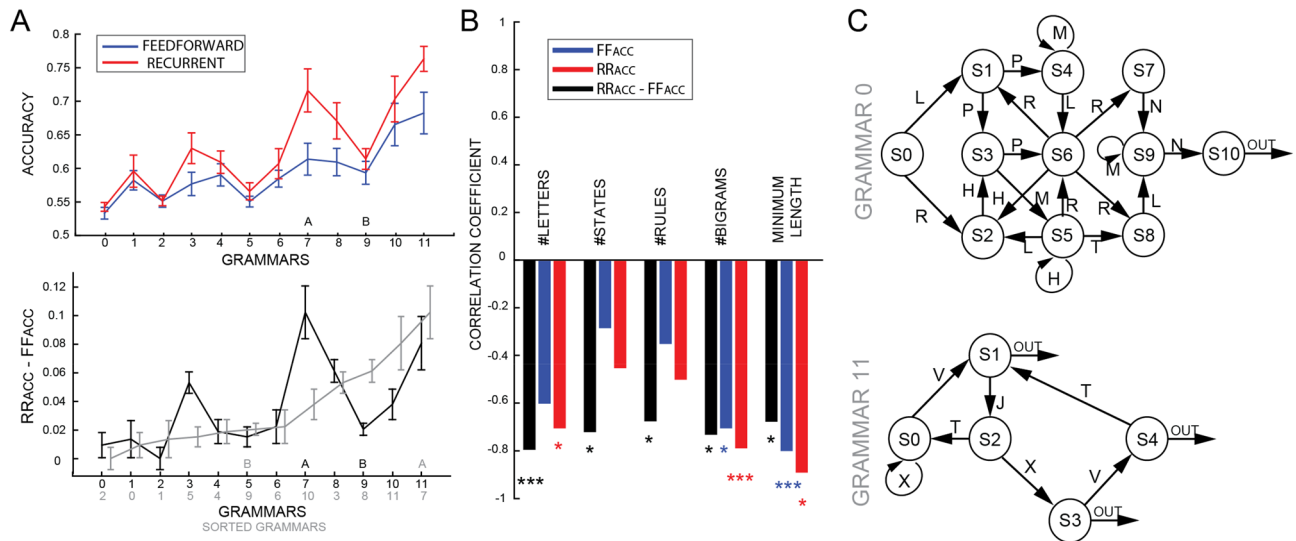
Figure 5. (**A**) The upper panel shows the performance of FF (in blue) and RR (in red) architectures for each grammar (from 1 to 10 of the x-axis), averaging over 20 initialization over a training set of 500 trials. Error bars represent standard errors of the mean. Grammars A and B are respectively 7 and 9. The difference in the performance between the two architectures is shown in the lower panel (the same sorted difference in light grey). (**B**) Pearson indexes obtained correlating the performance of FF and RR networks and their difference (respectively in blue, red and black) with the 5 complexity metrics across the 10 grammars. (**C**) The two novel grammars we designed at the 2 extremes of the complexity index to test our hypothesis.

more explicitly than complex grammars, we can consider our results as supporting the compelling hypothesis that explicit knowledge is best modeled by recurrent architectures.

Discussion

Summary of the results.   In this study we demonstrated that recurrent neural networks mimic human artificial grammar learning more closely than feedforward architectures, irrespective of the grammar's level within the Chomsky hierarchy. This result supports the hypothesis that recursion is key in cognitive processes such as language[63–66]. As already mentioned, previous studies showed that humans can learn to classify sequences as correct or not according to grammatical rules, and such knowledge appears to be to some extent implicit[22,28,67]. However, in a usual AGL experiment, participants are first asked to memorize a set of sequences (training phase) and then to classify a new set as correct or not (testing phase). Here, we combined the two phases such that training and testing occur at the same time, allowing us to track the learning dynamics as it progresses. This design let us compare the participants' learning with the artificial networks' one. Importantly, we showed that both feedforward and recurrent neural networks can learn artificial grammars within the same limited number of trials as for human participants. However, the overall behavior of the recurrent networks, and in particular their learning dynamics, was closer to human behavior.

Related work on artificial grammars and neural networks.   In our simulations, we contrasted the performance of feedforward and recurrent architectures in AGL tasks. The two architectures are substantially different, and represent distinct functional processes. Feedforward networks allow the information to flow one way from input to output. In feedforward architectures the sequence is processed at once, with a single sweep. A very similar approach is the competitive chunking network[21], in which the whole sequence is processed as one to identify hierarchical, nested patterns (i.e. the chunks). In this model several layers can be involved in the process before classification occurs[68]. However, our results suggest that FF networks having only 2 layers provide the best classification in AGL tasks within our experimental constraints (with a limited training dataset, comparable to human learning). Possibly, deeper networks would improve classification performance given a larger amount of trials, and may be more powerful with more complex grammars; however, we lack experimental studies showing how humans would learn these more complicated structures. On the other hand, recurrent models process the sequence differently, receiving one letter at a time and building short and long temporal dependencies between items[69]. The learning occurs by reinforcing associations that appear frequently within the training set, and it has been applied successfully to finite state grammars since the early '90s'[25,70,71]. Since these early studies investigating connectionist models of artificial grammar learning, recurrent neural networks have been proposed as an ideal candidate to investigate how learning occurs in artificial grammar experiments, particularly in the context of implicit learning[25,72]. In particular, two properties have been considered compelling in recurrent neural architectures: first, their sensitivity to temporal structure; second, the fact that the rules' representations are distributed over the weights of the networks, as assumed in implicit learning models[24,73]. Previous studies have investigated how neural networks learn to encode the states of Finite State Automata (FSA, equivalent to regular grammars in Chomsky's hierarchy), demonstrating that after learning each automaton (i.e. grammar) state is

Authors	Grammar	Network	Total training examples
Cleeremans et al.[25]	Regular	Recurrent	60,000
Servan-Schreiber et al.[72]	Regular	Recurrent	60,000
Cleeremans and McClelland[24]	Regular	Recurrent	60,000
Giles et al.[72]	Regular	Recurrent	150,000
Dienes[70]	Regular	Recurrent	20,000
Rodriguez et al.[79]	CFG	Recurrent	~ 15,000
Bodén and Wiles[98]	CFG/CSG	Recurrent	10,000
Kinder and Shanks[89]	Regular	Recurrent	2880
Gers and Schmidhuber[81]	CFG/CSG	Recurrent (LSTM)	10e3 to 10e7
Boucher and Dienes[23]	Regular	Recurrent and chunking model	~ 1000
Tunney and Shanks[99]	Regular	Recurrent	3200
Petersson et al.[78]	Regular	Recurrent	25,400
Wierzchon and Barbasz[82]	Regular	Feedforward	96,000
Kinder and Lotz[86]	Regular	Recurrent and chunking model	~ 1000
Duarte et al.[75]	Regular	Recurrent (spiking)	10e3 to 10e5
Cohen et al.[83]	Regular	Recurrent	225,000
Alamia et al.[100]	**Regular/CFG/CSG**	**Feedforward and recurrent**	**500**

Table 1. Representative overview of studies training artificial neural networks to classify novel strings. Most studies employed recurrent networks over regular grammars. The last column provides the total number of sequences the network was trained on (sequences per epoch, times the number of epochs). Our study, highlighted in bold, employs the smallest number of training examples, a few orders of magnitude smaller than most studies, and similar to typical experimental studies in humans.

represented in the activation of the recurrent layers[25]. Similarly to models trained via backpropagation, previous results demonstrated that neural networks whose learning is mediated by synaptic plasticity (e.g. STDP) can also succeed in sequence learning[74] and specifically in artificial grammars[75]. In Duarte, Series and Morrison[75], the authors demonstrated that a spiking network can learn a regular grammar by mere exposure to correct strings, being able to predict the probability distribution of a subsequent letter, and eventually to classify the legality of the whole string. Additionally, by considering the transmission delays in the communication between neurons it would be possible to determine how more biologically plausible models learn the grammars and possibly model response time in human subjects[76,77]. Subsequent work expanded this conclusion, demonstrating that second-order recurrent networks can infer the full FSA after an appropriate amount of training examples or epochs[71]. Context free and context specific grammars have also been studied by means of simple recurrent networks. One study proved that SRNs can learn to predict the next letter in a simple CF language, in which sequences are composed of two groups of letters ('A' and 'B') and organized according to the structure A^nB^n. The authors, interpreting recurrent networks as dynamical systems[78], demonstrated that the model learns to count items from each group in order to successfully determine the grammar's rules[79]. In a later study, they extended the results by showing that recurrent networks process and store information as if using a memory buffer, thereby solving both CF and CS grammars[80]. Another study investigated how Long Short Term Memory (LSTM) recurrent networks learn CF and CS grammars[81]. These LSTM architectures are becoming more prominent in natural language studies, given their ability to learn long-distance dependencies. Another interesting lien of research involves investigating how different architecture, like STDP, learn th grammars. However, in this and all previous studies, neural networks were trained on a very large number of examples over several epochs[75,82,83], as summarized in Table 1. This approach prevents a fair comparison between models and human performance, as human participants are able to classify novel strings above chance after only a few hundred examples.

A fair comparison between human and artificial neural networks. Table 1 provides a representative overview of studies investigating the behavior of artificial neural networks in classifying sequences from different grammars. First, it appears that previous studies have investigated mostly recurrent rather than feed-forward networks, even though implicit learning has been frequently related to feedforward processes[46,84,85] (see also next "Computational correlates of implicit and explicit processes" section). Second, all but two studies[23,86] employed thousands (or tens of thousands) of sequences to train the models' parameters (except Boucher & Dienes and Kinder & Lotz which used ~ 1000 examples across the training epochs). In our study, highlighted in yellow, we implemented a fairer comparison with human performance, as both artificial networks and human participants were trained on the same limited number of examples (~ 500). Interestingly, when trained over larger dataset (~ 10,000 sequences) we observed that recurrent network still perform better than feedforward ones on regular grammars (on both GA and GB, FF ~ 0.85%, RR ~ 0.95%—not shown), but the difference reduces in our Context-Free and Context-Sensitive grammars (both architectures on both grammars ~ 75%—not shown). In addition, differently than previous studies on artificial grammar learning in humans, we adopted an experimental design in which training and testing occurred at the same time, allowing a direct comparison of learning dynamics (i.e., learning curves) between humans and artificial neural networks. Previous studies

had compared recurrent networks and a chunking model (similar to a feedforward architecture, see paragraph "Related work on artificial grammars and neural networks") with human behavior, but with opposite results[23,86]. Our approach allows us to directly compare human performance with both feedforward and recurrent models, avoiding the limitations discussed above. All in all, this comparison reveals that recurrent models perform closer to humans than feedforward ones, except in more complicated—and supposedly implicit—grammars. For those grammars (e.g. Grammar B), human performance remains poor, and can equally well be accounted for by recurrent or feedforward models. Altogether, it is possible to speculate that the similarity between recurrent networks and neuronal dynamics in the brain may be an intriguing explanation for the results of our comparison[87,88], even though additional neuroimaging studies will be required to shed light on this interpretation. Moreover, another consideration that emerges from our comparison regards the crucial role of the learning rate in recurrent networks. Interestingly, in an attempt to model the behavior of Amnesic and Healthy subjects during an artificial grammar learning task, some authors reported a steady decrease in the classification performance of a simple recurrent network with the increase of its learning rate[89]—as if Amnesic patients had different learning rates compared to Healthy subjects. Their conclusions are in line with our observations, as they confirm that recurrent networks perform best in classifying novel sequences with low learning rate values. Finally, another appealing result of our investigation concerns the functional distinction between implicit and explicit processes in artificial grammar learning, as examined in the next section.

Computational correlates of implicit and explicit processes. In the two regular grammars we observed a significant difference in participants' awareness of the rules. Even though the task required participants to be aware of their learning process (i.e. intentional learning tasks), most participants failed in reporting the content of all or part of their acquired knowledge, showing different degree of awareness of the rules in each grammar. In grammar A participants performed better at the questionnaire than in grammar B, coherently with the hypothesis that simpler grammars are more likely to be learnt consciously[37,38,90]. The distinction between implicit and explicit processes has been expressly designed in an integrated model tested also on artificial grammar tasks[91,92], providing evidence in favor of the hypothesis that implicit processes precede explicit ones[93,94] and are prominently involved in complex grammars[37,38]. Interestingly, our results reveal that in the more implicit grammar B, both ANN architectures reliably tracked human behavior, whereas only recurrent networks achieved this goal in more explicit grammars. This result draws a compelling parallel between feedforward/recurrent and implicit/explicit processes, consistently with results in visual perception[46,95] and neuroscience[84,85].

Conclusion

Which neural network architecture matches human behavior in an artificial grammar learning task? We demonstrate in this study that recurrent neural network models are closer to humans than feedforward ones, irrespective of the grammars' level in the Chomsky's hierarchy. This result endorses recurrent models as the more plausible cognitive architecture underlying language, pointing to an essential role of recursion in grammar learning, even if additional neuroimaging studies will be required to investigate the underlying brain dynamics[96]. Moreover, our comparison between architectures over 10 different regular grammars demonstrates that recurrent models can better learn simpler grammars, which are also more likely to be learnt explicitly than complex grammars (as confirmed by our results on human participants). These results support the compelling hypothesis that explicit knowledge is best modeled by recurrent architectures. However, it could be fetching to generalize this conclusion to natural language acquisition[10,64,87,88]: the Chomsky hierarchy, despite being an excellent reference, does not embrace all natural grammars[12], and does not fully reflect human cognitive complexity[13]. Additional studies will investigate more ecological grammars to further corroborate our conclusion[97].

References

1. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. VanRullen, R. Perception science in the age of deep neural networks. *Front. Psychol.* **8**, 142. https://doi.org/10.3389/fpsyg.2017.00142 (2017).
3. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing [Review Article]. *IEEE Comput. Intell. Mag.* **13**, 55–75 (2018).
4. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
5. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
6. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **2**, 113–124 (1956).
7. Turing, A. M. On computable numbers, with application to the Entscheidungs problem. *Proc. Lond. Math. Soc.* https://doi.org/10.1112/plms/s2-42.1.23 (1937).
8. Post, E. Recursively enummerable sets of postive integers and their decision problems. *ams.org* (1944).
9. Chomsky, N. On certain formal properties of grammars. *Inf. Control* **2**, 137–167 (1959).
10. Tecumsehfitch, W. & Friederici, A. D. Artificial grammar learning meets formal language theory: An overview. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1933–1955 (2012).
11. Joshi, A. K. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?. *Nat Lang. Parsing: Psychol. Comput. Theor. Perspect.* https://doi.org/10.1017/cbo9780511597855.007 (1985).
12. Jäger, G. & Rogers, J. Formal language theory: Refining the Chomsky hierarchy. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1956–1970 (2012).

13. Öttl, B., Jäger, G. & Kaup, B. Does formal complexity reflect cognitive complexity? Investigating aspects of the Chomsky hierarchy in an artificial language learning study. *PLoS ONE* **10**, e0123059 (2015).
14. Pothos, E. M. Theories of artificial grammar learning. *Psychol. Bull.* **133**, 227–244 (2007).
15. Reber, A. S. Implicit learning of artificial grammars. *J. Verbal Learn. Verbal Behav.* **6**, 855–863 (1967).
16. Shanks, D. & Stjohn, M. Characteristics of dissociable human learning-systems. *Behav. Brain Sci.* **17**(3), 367–395. https://doi.org/10.1017/S0140525X00035032 (1994).
17. Brooks, L. R. & Vokey, J. R. Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *J. Exp. Psychol. Gen.* **120**, 316–323 (1991).
18. Vokey, J. R. & Brooks, L. R. Salience of item knowledge in learning artificial grammars. *J. Exp. Psychol. Learn. Mem. Cogn.* **18**, 328–344 (1992).
19. Dulany, D. E., Carlson, R. A. & Dewey, G. I. A case of syntactical learning and judgment: How conscious and how abstract?. *J. Exp. Psychol. Gen.* **113**, 541–555 (1984).
20. Reber, A. S. & Lewis, S. Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition* **5**, 333–361 (1977).
21. Servan-Schreiber, E. & Anderson, J. R. Learning artificial grammars with competitive chunking. *J. Exp. Psychol. Learn. Mem. Cogn.* **16**, 592–608 (1990).
22. Knowlton, B. J. & Squire, L. R. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 169–181 (1996).
23. Boucher, L. & Dienes, Z. Two ways of learning associations. *Cogn. Sci.* **27**, 807–842 (2003).
24. Cleeremans, A. & McClelland, J. L. Learning the structure of event sequences. *J. Exp. Psychol. Gen.* **120**, 235–253 (1991).
25. Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L. Finite state automata and simple recurrent networks. *Neural Comput.* **1**, 372–381 (1989).
26. Reber, A. S. Implicit learning and tacit knowledge: An essay on the cognitive unconscious. *Implicit Learn. Tacit Knowl. Essay Cogn. Unconscious* https://doi.org/10.1093/acprof:oso/9780195106589.001.0001 (2008).
27. Cleeremans, A., Destrebecqz, A. & Boyer, M. Implicit learning: News from the front. *Trends Cogn. Sci.* **2**, 406–416 (1998).
28. Reber, A. S. Implicit learning of artificial grammars. (1967).
29. Squire, L. R. & Dede, A. J. O. Conscious and unconscious memory systems. *Cold Spring Harb. Perspect. Biol.* **7**, a021667 (2015).
30. Berry, D. C. & Broadbent, D. E. On the relationship between task performance and verbal knowledge. *Q. J. Exp. Psychol.* **36A**(209–231), 209–231 (1984).
31. Berry, D. C. & Broadbent, D. E. In *Complex Problem Solving: The European Perspective.* 131–150 (1995).
32. Newell, B. R. & Shanks, D. R. Unconscious influences on decision making: A critical review. *Behav. Brain Sci.* **37**, 1–19 (2014).
33. Alamia, A. *et al.* Unconscious associative learning with conscious cues. *Neurosci. Conscious.* **2016**, 1–10 (2016).
34. Cleeremans, A. Computational correlates of consciousness. *Prog. Brain Res.* **150**, 81–98 (2005).
35. Koch, K. In *The Quest for Consciousness—A Neurobiological Approach* 221–233 (2004).
36. Reber, A. S. Implicit learning and tacit knowledge. *J. Exp. Psychol. Gen.* **118**, 219–235 (1989).
37. Reber, A. S. Implicit learning of synthetic languages: The role of instructional set. *J. Exp. Psychol. Hum. Learn. Mem.* **2**, 88–94 (1976).
38. Halford, G. S., Wilson, W. H. & Phillips, S. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* **21**, 803–864 (1998).
39. Capacity, H. P. The magic number seven—The limits of human cognition. *Stress Int. J. Biol. Stress* https://doi.org/10.1242/jeb.110312 (2007).
40. Cowan, N. The magical mystery four: How is working memory capacity limited, and why?. *Curr. Dir. Psychol. Sci.* **19**, 51–57 (2010).
41. Rah, S. K. Y., Reber, A. S. & Hsiao, A. T. Another wrinkle on the dual-task SRT experiment: It's probably not dual task. *Psychon. Bull. Rev.* **7**, 309–313 (2000).
42. Röttger, E., Haider, H., Zhao, F. & Gaschler, R. Implicit sequence learning despite multitasking: The role of across-task predictability. *Psychol. Res.* https://doi.org/10.1007/s00426-017-0920-4 (2017).
43. Smith, J. G. & McDowall, J. The implicit sequence learning deficit in patients with Parkinson's disease: A matter of impaired sequence integration?. *Neuropsychologia* **44**, 275–288 (2006).
44. Dienes, Z. & Berry, D. C. Implicit learning: Below the subjective threshold?. *Psychon. Bull. Rev.* **4**, 3–23 (1997).
45. Dienes, Z. & Perner, J. A theory of implicit and explicit knowledge. *Behav. Brain Sci.* **22**, 735–808 (1999).
46. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
47. Supèr, H., Spekreijse, H. & Lamme, V. A. F. Two distinct modes of sensory processing observed in monkey primary visual cortex (VI). *Nat. Neurosci.* **4**, 304–310 (2001).
48. Chollet, F. Keras: Deep Learning library for Theano and TensorFlow. *GitHub Repos.* 1–21 (2015). https://doi.org/10.1111/j.1439-0310.1985.tb00118.x
49. GoogleResearch. TensorFlow: Large-scale machine learning on heterogeneous systems. *Google Res.* (2015). https://doi.org/10.1207/s15326985ep4001
50. Robbins, H. & Monro, S. In *Herbert Robbins Selected Papers* 102–109 (1985). https://doi.org/10.1007/978-1-4612-5110-1_9
51. Nesterov, Y. A method of solving a convex programming problem with convergence rate O (1/k2). *Sov. Math. Dokl.* **27**, 372–376 (1983).
52. Mozer, M. C. A focused backpropagation algorithm for temporal pattern recognition. *Complex Syst.* **3**, 349–381 (1989).
53. Werbos, P. J. Backpropagation through time: What it does and how it does. *Proc. IEEE* **78**, 1550–1561 (1990).
54. Tieleman, T., Hinton, G. E., Srivastava, N. & Swersky, K. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* **4**, 26–31 (2012).
55. JASP Team. JASP (Version 0.8.6.0). *[Computer software].* (2018).
56. Love, J. *et al.* Software to sharpen your stats. *APS Obs.* **28**, 27–29 (2015).
57. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
58. Masson, M. E. J. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behav. Res. Methods* **43**, 679–690. https://doi.org/10.3758/s13428-010-0049-5 (2011).
59. Bernardo, J. M. & Smith, A. F. M. *Bayesian Theory.* (2008). https://doi.org/10.1002/9780470316870.
60. Cadogan, A. Interpretation of sensitivity and specificity. *N. Z. J. Sport. Med.* (2012).
61. Cleeremans, A. & Dienes, Z. In *The Cambridge Handbook of Computational Psychology* 396–421 (2012). https://doi.org/10.1017/cbo9780511816772.018
62. Katan, P. & Schiff, R. Does complexity matter? Meta-Analysis of learner performance in artificial grammar tasks. *Front. Psychol.* **5**, 1084 (2014).
63. Steedman, M. Plans, affordances, and combinatory grammar. *Linguist. Philos.* **25**, 723–753. https://doi.org/10.1023/A:1020820000972 (2002).
64. Jackendoff, R. What is the human language faculty?: Two views. *Language* **87**, 586–624 (2011).
65. Jackendoff, R. Parallels and nonparallels between language and music. *Music Percept. Interdiscip. J.* **26**, 195–204 (2009).

66. Fischmeister, F. P., Martins, M. J. D., Beisteiner, R. & Fitch, W. T. Self-similarity and recursion as default modes in human cognition. *Cortex* **97**, 183–201 (2017).
67. Rohrmeier, M., Fu, Q. & Dienes, Z. Implicit learning of recursive context-free grammars. *PLoS ONE* **7**, e45885 (2012).
68. Buchner, A. Indirect effects of synthetic grammar learning in an identification task. *J. Exp. Psychol. Learn. Mem. Cogn.* **20**, 550–566 (1994).
69. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
70. Dienes, Z. Connectionist and memory-array models of artificial grammar learning. *Cogn. Sci.* **16**, 41–79 (1992).
71. Giles, C. L. *et al.* Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Comput.* **4**, 393–405 (1992).
72. Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Mach. Learn.* **7**, 161–193 (1991).
73. Cleeremans, A. & Dienes, Z. Computational models of implicit learning. In *The Cambridge handbook of computational psychology* ed R. Sun (Cambridge University Press, London, UK, pp. 396-421, 2008).
74. Wang, Q., Rothkopf, C. A. & Triesch, J. A model of human motor sequence learning explains facilitation and interference effects based on spike-timing dependent plasticity. *PLoS Comput. Biol.* **13**, e1005632 (2017).
75. Duarte, R., Seriès, P. & Morrison, A. Self-Organized Artificial Grammar Learning in Spiking Neural Networks. In *Proc. 36th Annu. Conf. Cogn. Sci. Soc.* 427–432 (2014).
76. Madadi Asl, M., Valizadeh, A. & Tass, P. A. Dendritic and axonal propagation delays determine emergent structures of neuronal networks with plastic synapses. *Sci. Rep.* **7**, 39682 (2017).
77. Madadi Asl, M., Valizadeh, A. & Tass, P. A. Delay-induced multistability and loop formation in neuronal networks with spike-timing-dependent plasticity. *Sci. Rep.* **8**, 1–5 (2018).
78. Petersson, K. M., Grenholm, P. & Forkstam, C. Artificial grammar learning and neural networks. In *Proc. 27th Annu. Conf. Cogn. Sci. Soc.* 1726–1731 (2005).
79. Rodriguez, P., Wiles, J & Elman, J. L. A recurrent neural network that learns to count. *Connect. Sci.* **11**, 5–40 (1999).
80. Rodriguez, P. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Comput.* **13**, 2093–2118 (2001).
81. Gers, F. A. & Schmidhuber, J. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Netw.* **12**, 1333–1340 (2001).
82. Wierzchon, M. & Barbasz, J. A Connectionist model of artificial grammar learning: simulations based on Higham (1997) indexes of knowledge representation. In *29th Annu. Conf. Cogn. Sci. Soc. Washington, DC* (2008).
83. Cohen, M., Caciularu, A., Rejwan, I. & Berant, J. Inducing Regular Grammars Using Recurrent Neural Networks. 1–5 (2017).
84. Boly, M. *et al.* Preserved feedforward but impaired top-down processes in the vegetative state. *Science* **332**, 858–862 (2011).
85. Koch, C., Massimini, M., Boly, M. & Tononi, G. Neural correlates of consciousness: Progress and problems. *Nat. Rev. Neurosci.* **17**, 395–395 (2016).
86. Kinder, A. & Lotz, A. Connectionist models of artificial grammar learning: What type of knowledge is acquired?. *Psychol. Res.* **73**, 659–673 (2009).
87. Corballis, M. C. In *Language and Recursion*, Vol. 9781461494, 27–36 (2014).
88. Vergauwen, R. In *Language and Recursion*, Vol. 9781461494, 169–179 (2014).
89. Kinder, A. & Shanks, D. R. Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *J. Cogn. Neurosci.* **13**, 648–669 (2001).
90. Sun, R. & Peterson, T. Learning in reactive sequential decision tasks: the CLARION model. In *Proceedings of International Conference on Neural Networks (ICNN'96)* Vol. 2, 1073–1078 (1994).
91. Sun, R. The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cogn. MultiAgent Interact.* https://doi.org/10.1017/CBO9780511610721.005 (2006).
92. Sun, R., Zhang, X., Slusarz, P. & Mathews, R. The interaction of implicit learning, explicit hypothesis testing learning and implicit-to-explicit knowledge extraction. *Neural Netw.* **20**, 34–47 (2007).
93. Cleeremans, A. Principles for Implicit Learning. (1997).
94. Windey, B. & Cleeremans, A. Consciousness as a graded and an all-or-none phenomenon: A conceptual analysis. *Conscious. Cogn.* https://doi.org/10.1016/j.concog.2015.03.002 (2015).
95. VanRullen, R. & Thorpe, S. J. The time course of visual processing: From early perception to decision-making. *J. Cogn. Neurosci.* **13**, 454–461 (2001).
96. Rohrmeier, M., Dienes, Z., Guo, X. & Fu, Q. In *Language and Recursion*, Vol. 9781461494, 67–85 (2014).
97. Smith, A. Grammar inference using recurrent neural networks. (Dep. Comput. Sci. Univ. San., 2003).
98. Bodén, M., & Wiles, J. Context-free and context-sensitive dynamics in recurrent neural networks. *Connect. Sci.* https://doi.org/10.1080/095400900750060122 (2000).
99. Tunney, R.J., & Shanks, D.R. Subjective measures of awareness and implicit cognition. *Mem. Cognit.* **31**, 1060–1071. https://doi.org/10.3758/BF03196127 (2003).
100. Alamia, A., Gauducheau, V., Paisios, D., & VanRullen, R. "Which Neural Network Architecture matches Human Behavior in Artificial Grammar" Learning? arXiv:1902.04861 (2019).

Acknowledgements

Author contributions

Experimental design: A.A., R.V.R. Data collection: A.A., V.G., D.P. Data analysis: A.A., V.G., D.P. Funding acquisition: R.V.R. Supervision: R.V.R., A.A. Writing—original draft: A.A. Writing—review and editing: R.V.R., V.G., D.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79127-y.

Correspondence and requests for materials should be addressed to A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.