Rapid Learning of Temporal Dependencies at Multiple Timescales

Cybelle M. Smith, Sharon L. Thompson-Schill, and Anna C. Schapiro

Abstract

■ Our environment contains temporal information unfolding simultaneously at multiple timescales. How do we learn and represent these dynamic and overlapping information streams? We investigated these processes in a statistical learning paradigm with simultaneous short and long timescale contingencies. Human participants ($n = 96$) played a game where they learned to quickly click on a target image when it appeared in one of nine locations, in eight different contexts. Across contexts, we manipulated the order of target locations: at a short timescale, the order of pairs of sequential locations in which the target appeared; at a longer timescale, the set of locations that appeared in the first versus the second half of the game. Participants periodically predicted the upcoming target location, and later performed similarity judgments comparing the games based on their order properties. Participants showed context-dependent sensitivity to order information at both short and long timescales, with evidence of stronger learning for short timescales. We modeled the learning paradigm using a gated recurrent network trained to make immediate predictions, which demonstrated multilevel learning timecourses and patterns of sensitivity to the similarity structure of the games that mirrored human participants. The model grouped games with matching rule structure and dissociated games based on low-level order information more so than high-level order information. The work shows how humans and models can rapidly and concurrently acquire order information at different timescales. ■

INTRODUCTION

The environment contains temporal dependencies that unfold simultaneously at multiple timescales. For example, a baseball fan can anticipate the trajectory of a just-pitched ball as well as the switching of jerseys in the field between the top and bottom half of each inning. To fully represent, simulate, and anticipate changes in the environment, humans must learn these concurrent temporal dependencies. Learning at multiple timescales has been documented in a variety of domains, including language (Saffran & Wilson, 2003), event segmentation (Shin & DuBrow, 2021; Davachi & DuBrow, 2015), planning (Momennejad, 2024; Lee, Aly, & Baldassano, 2021), motor learning (Krakauer, Hadjiosif, Xu, Wong, & Haith, 2019), and visual statistical learning (Karuza, Kahn, Thompson-Schill, & Bassett, 2017; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Our study adds to this literature by testing whether humans can rapidly learn simultaneous regularities occurring at multiple timescales. In particular, we explore how sensitive humans as well as neural network models are to slow background statistical dependencies when they are focused on making short-term predictions in the immediate task at hand.

Studies on statistical learning of auditory content (language and music) suggest that multiple timescales of temporal statistics can be rapidly acquired under at least some circumstances. For example, infants can learn transition probabilities among both syllables (Saffran, Aslin, & Newport, 1996) and pairs of syllables ("words"; Saffran & Wilson, 2003) after only a few minutes of exposure. Moreover, infants and adults are sensitive to nonadjacent dependencies among nonsense words within a single learning session (Misyak, Christiansen, & Tomblin, 2010; Gómez, 2002), and adults can learn higher order statistics among tones after a mere 5 min of exposure (Furl et al., 2011). In these cases, higher order learning is often modulated by the salience of embedded fast temporal statistics. For example, the presence of high adjacent transition probabilities impedes acquisition of nonadjacent dependencies (Gómez, 2002), raising the possibility of attentional trade-offs for statistical learning at different temporal scales. Indeed, when intervening sounds are of a different type from those related through a nonadjacent dependency, the dependency becomes easier to learn (Creel, Newport, & Aslin, 2004; Newport & Aslin, 2004). It is unknown whether these differences are needed when the higher order dependencies unfold at a slower timescale (more than a few seconds), where they may be less likely to be confused with the short timescale information.

In the visual, spatial, and motor domains, humans rapidly learn short temporal dependencies across a variety of paradigms in under an hour of exposure (Krakauer et al., 2019; Fiser & Aslin, 2002). However, many paradigms showing sensitivity to longer timescale statistics appear

University of Pennsylvania

Journal of Cognitive Neuroscience 36:11, pp. 2343–2356
https://doi.org/10.1162/jocn_a_02232

to involve extensive training. For example, given many sessions of training, participants in visuospatial search and visuomotor learning tasks are able to use temporal context going back at least three trials to anticipate the upcoming target location (Cleeremans & McClelland, 1991; Lewicki, Czyzewska, & Hoffman, 1987). Similarly, in the motor learning literature, after extended training, shuffling previously learned motor chunks results in a decrement in performance relative to intact sequences, consistent with learning of higher-level order among the chunks (Sakai, Kitaguchi, & Hikosaka, 2003). It is unclear whether there are contexts in which simple but slow visuospatial or motor statistics can be acquired rapidly, and whether the presence of faster temporal dependencies interferes with learning of slower visuospatial or motor regularities.

We report the results of a preregistered study in which humans learned nested temporal dependencies at fast and slow timescales in a visuo-spatial motor learning task inspired by the carnival game "whack-a-mole." Participants played different mini-games of whack-a-mole in which the temporal dependencies among target locations on different trials varied. Participants were periodically asked to predict the upcoming target location. They were also asked to judge the similarity of the different mini-games on the basis of their temporal structure (both fast and slow). Using the results from both the prediction task and the similarity judgment task, we were able to track participant learning over time and assess whether both slow and fast timescale dependencies were being learned within a single session. We predicted on the basis of our pilot data that participants would be able to learn both timescales of dependencies, with an advantage for the more immediate fast timescale.

We simulated performance on our behavioral tasks (online predictions and posttraining similarity judgments) using a gated recurrent neural network with a single hidden layer. We trained the model to predict the subsequent target location, using input stimuli generated in the same way as our human behavioral experiment, and evaluated to what extent the model became sensitive to longer timescale dependencies over training, and its degree of match to the human behavioral data.

METHODS

Human Behavioral Methods

Participants

One hundred three participants were recruited for course credit or monetary compensation from the University of Pennsylvania participant pool. We obtained informed consent from all participants in accordance with the University of Pennsylvania institutional review board. Participants were native English speakers with normal/corrected-to-normal vision and hearing. We used preregistered exclusion criteria as follows: (1) similarity judgment task attention check accuracy below 70%; (2) during gameplay, missing more than 25% of responses; (3) during gameplay, not responding on any trial for two consecutive games or more. Following these criteria, seven participants were excluded and replaced, leaving our final predetermined sample size of 96 participants (aged 18 to 59 years, mean = 25.2 years, 14 participants were over age 30 years, five participants did not disclose their age). Our target sample size was determined by power analysis of a smaller pilot experiment ($n = 32$) using measures derived from the similarity judgment task described further below, with the simr package *powerSim*, indicating that power was over 95% for all key contrasts (without correction for multiple comparisons). Preregistered methods are available at https://archive.org/details/osf-registrations-egukx-v1.

Materials

Participants played eight mini-games of whack-a-mole. Each mini-game contained a distinctive background image (similar to a game board or arena) and a thematically related target image (always an animate object; Figure 1). Both the background and target images were in cartoon style. Background images were all covered with gridlines of similar size, and a gray semitransparent overlay was used to delineate the arena during gameplay. The following background and target images were used for the eight minigames: (1) background: desert with skulls carved into a plateau, target: skull; (2) background: forest temple, target: fairy; (3) background: desert with pyramid structure and pond, target: camel; (4) background: forest with blossoming cherry trees, target: animate cherry blossom with face; (5) background: snowy temple with bridge and large statues, target: yeti; (6) background: terraced land with large mushrooms, target: animate mushroom with face; (7) background: underground river and rock formation with crystals, target: bat; (8) background: medieval encampment, target: dragon. The order of locations in which the target image appeared in each game was counter-balanced across participants, such that no background/target image pair was consistently associated with any set of order rules. Participants were given an additional practice game in which the target appeared in completely random locations. The background and target images used for this game were distinct from those used during training, and consisted of a swimming pool and a goldfish.

Procedure

Exposure to games. Participants were informed that they would play a series of mini-games similar to whack-a-mole. They were asked to pay attention to how similar the mini-games were to each other and informed that they would answer questions comparing the games afterward. They were particularly asked to pay attention to the order in which the target appeared during the games, and differences in that order among the games. Participants played

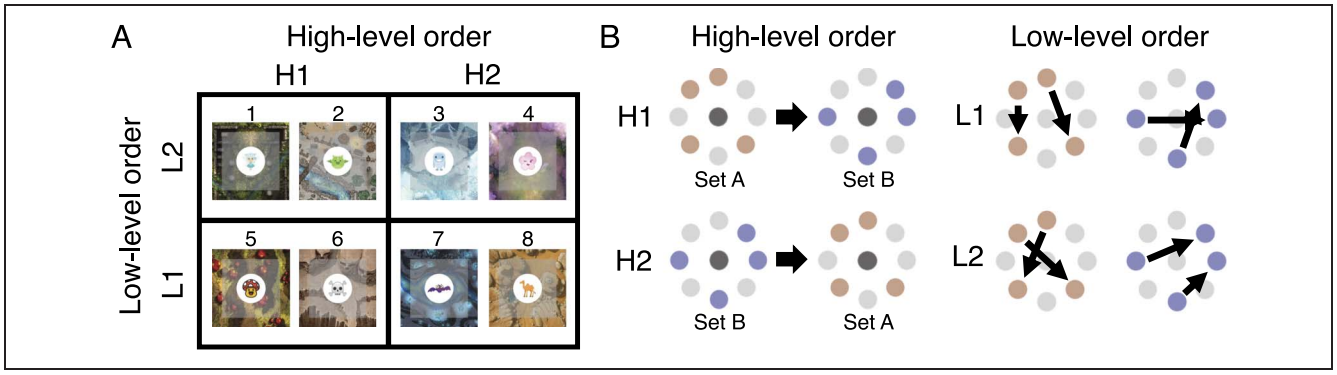

Figure 1. Experiment design. (A) High- and low-level order rules were manipulated within participants in a 2 × 2 design. (B) High-level order determined the set of target locations encountered in the first versus second half of the game. Low-level order determined which paired associations were encountered among the target locations on adjacent trials during the game.

each of eight mini-games once per round in a series of eight rounds. Participants received points for their performance and could view the point total for each game during that game, and their overall point total after each game. Each mini-game began with a slide that introduced the background of the game (e.g., a drawing of a snowy mountain temple scene with gridlines over it) and the target image that they needed to click (e.g., a cartoon yeti), which was superimposed on the background. During each mini-game, a board was displayed with nine dark circles overlaid on it that resembled "holes": one center hole and eight holes evenly spaced on the periphery in a circular arrangement. During each trial of the game, a target image appeared in one of the nine holes and the participant had to click on the target before it disappeared to receive 25 points. The time limit to receive points was 800 msec after target onset. If they did not click the target in time, they heard a laughing noise and did not receive any points. Participants were informed before playing the games that the laughing noise indicated that they were too slow. Clicking the target ended the trial. If the participant did not click the location that the target appeared in, the trial ended 2200 msec after target offset. The intertrial interval was 500 msec (from the end of one trial to the appearance of the target in the next trial).

Game structure. Each game lasted 27 trials. Games were divided in a 2 × 2 design by adopting one of two high-level order rules (slow timescale) and one of two low-level order rules (fast timescale; Figure 1). Two of the eight games were assigned to each combination of high- and low-level order rules. The eight peripheral (noncenter) target locations were randomly partitioned into two sets of four peripheral locations. High-level order rules governed which set (A or B) was used for target locations in the first half of the game; in the second half of the game, the target appeared in the remaining set of locations. Locations were also grouped into ordered pairs such that if one peripheral location was visited on a given trial, the target would appear in the second location on the following trial 100% of the time. Low-level order rules governed the second location in each of the four ordered pairs of locations. The second location of each pair was swapped across games with different low-level order. Before presenting each ordered pair of locations, the target appeared in the center. By separating pairs of locations with interleaved center location trials, we ensured that participants were not able to rely solely on the current stimulus to predict the first item in each pair, and must keep track of temporal information. In the middle of each game, the target appeared in the center three times in a row to help form an event boundary separating the first and second half of the game. Thus, although only eight pairs of peripheral locations were presented in each game (16 trials), 27 trials were presented in each game including the center location.

All participants were assigned to a unique randomized list of games, belonging to one of 16 counterbalancing groups. Counterbalancing groups were designed such that each game stimulus was assigned to each of the four joint order conditions (H1L1, H1L2, H2L1, H2L2) an equal number of times across lists. Each game stimulus shared its high-level or low-level order condition with each other game stimulus an equal number of times across the counterbalancing. The order of presentation of paired associate locations within each half of the game was randomized. Two pairs were presented twice in each half of gameplay (the remaining two pairs were presented twice in the second half). The games were played in "rounds" of eight such that each mini-game was played once per round. The order of the games within a round was randomized.

Prediction probe trials. During gameplay, participants were periodically probed on their order knowledge by explicitly predicting the upcoming target location. During probe trials, the eight peripheral locations were covered by a "?" and the participant had to select where they thought the target would appear next (barring the center hole). Responses were self-paced. Each probe trial was followed by a 1000-msec delay and then by an ordinary game trial that provided participants with feedback on

their predictions. There was one prediction probe trial each time a game was played, except for the final round, when there were three probe trials per game. Prediction probe trials occurred throughout the game at all noncenter positions but were more likely to appear as the second trial in the game (37.5% of the time and then roughly evenly distributed across the remaining trial numbers and positions). Participants received 750 bonus points each time they selected the actual target location. Depending on when the trial appeared within the game, the maximum possible accuracy was either 50% (for the first location in an ordered pair) or 100% (for the second location in an ordered pair).

Similarity judgment task. After playing the games, participants judged the similarity of the games in a two-alternative forced-choice task. Participants were presented with a picture representing one game at the top and were asked which of two games presented at the bottom was more similar to the game at the top. Figure 1A displays the full set of stimuli presented to represent the games. All participants were asked to base their judgments on the order of locations in the games. They were randomly assigned to one of two conditions for finer grained instructional wording (48 participants per condition): In Condition 1, they were told to base similarity on "when/where the target appeared," and in Condition 2, they were told to base similarity on the "sequence of locations [the] target appeared in." We predicted based on initial pilot data that instructions condition would modulate sensitivity to low-level order on the similarity judgment task; specifically, we expected greater sensitivity to low-level order in instructions Condition 2.

Similarity judgments were performed on distinct trial types: (a) Attention check. One choice was the same game as the comparison game (with the same target image). This trial type was used for performance-based exclusion. Example (using game ID numbers from Figure 1A): target = 1, left = 1, right = 5. The image on the left is the same as the target image, so choose left. In all examples below, the left choice would be the more correct one (except (g)). (b) Same rules (but different game identity / target image) versus game with different high- and low-level order condition. Performance on this trial type may reflect sensitivity to either low- or high-level order information. Example: target = 1, left = 2, right = 8. (c) Same rules versus game with different low-level order condition. Above-chance accuracy would reflect sensitivity to low-level order information. Example: target = 1, left = 2, right = 5. (d) Same rules versus game with different high-level order condition. Above-chance accuracy would reflect sensitivity to high-level order information. Example: target = 1, left = 2, right = 4. (e) Both low- and high-level order differ versus only high-level order differs from the comparison game. Above-chance accuracy would reflect sensitivity to low-level order information. Example: target = 1, left = 4, right = 8. (f) Both low- and high-level order differ versus only low-level order differs from the comparison game. Above-chance accuracy would reflect sensitivity to high-level order information. Example: target = 1, left = 5, right = 8. (g) One game matches the comparison game on low-level order, and the other matches the comparison game on high-level order. This trial type assesses bias toward relying on high- versus low-level order information when making similarity judgments. Example: target = 1, left = 4, right = 5.

Analysis

Logistic regression models were fit with the *glmer* function using the *lme4* package in R. Nested model comparisons with the *anova* function were used to test for significant effects. The *emmeans* package was used for statistical reporting of condition means. Reported p values are uncorrected except where specified.

Online prediction judgments. Following our preregistered analysis plan, we fit logistic regression models to assess indices of sensitivity to high- and low-level order information and their change over time. Low-level sensitivity was assessed using responses to probes at the second location among paired locations. It was computed as: [Proportion congruent with low-level order rule] – [Proportion congruent with opposite low-level order rule]. High-level sensitivity was assessed using responses to probes at the first location among paired locations. It was computed as: [Proportion congruent with high-level order rule] – [Proportion congruent with opposite high-level order rule]. This was equivalent to subtracting the proportion of trials with responses drawn from the incorrect set of four locations, given the game rules and game half, from the correct set of four locations. High-level sensitivity measures were also computed separately using only the first noncenter trial of each game. Responses to the first noncenter trial in each game indexed sensitivity of location predictions to the target and context image. Our full model used the following formula in R notation:

```
model = glmer(prediction_congruent_with_rule
              ~ time + (1|participant) + (1|game_id),
              family = "binomial")
```

Here, time was indexed as the number of individual 27-trial games played since the beginning of the session, divided by 100 to aid with parameter estimation. We included crossed random intercepts for participant and game identity (which of the eight games was being played, in terms of the target and context image; because game positions were assigned to random locations for each unique participant, there was no reason to anticipate consistent effects of rule condition across participants). Separate models were fit for modeling sensitivity to the high- and low-level rules. Reduced models excluding time as a variable were used to assess mean performance

across the entire session. Nested model comparisons were used to assess the presence of an interaction between sensitivity and time.

RTs. We also examined RTs during ordinary trials that did not contain prediction probes. Given the large number of possible target locations, however, we found that it was difficult to distinguish between RT changes driven by overall familiarity with the task/game structure and context-specific sensitivity to low- and high-level order rules. We thus focused on accuracy measures in our analyses.

Similarity judgments. We fit a logistic regression model to the similarity judgment data, including trials that tested high-level order knowledge, low-level order knowledge, or both, with the following formula in R notation:

```
model = glmer(correct ~ instructions
        * trial_type_indices + (1|item) + (1|subj),
        family = "binominal")
```

The variable trial_type_indices was a factor with three levels coding whether a trial type was an index of sensitivity to high-level order, sensitivity to low-level order, or sensitivity to both; thus, the high- and low-level order levels were each aggregated across two different original trial types, in which one option matched the target on either both or neither dimension. We tested for the significance of fixed effects of interest using the *emmeans* package and nested model comparisons. We also ran a second logistic regression model and evaluated its output in the same way, but with disaggregated trial types (i.e., a five-factor model of trial type with a factor for each of trial types b–f above). We also separately examined performance on Trial Types (a) (attention check) and (g) (pitting high- and low-level order information against each other).

Modeling Methods

Architecture

A gated recurrent unit model with a single hidden layer (n = 48 model instances) was trained to predict the upcoming location in a whack-a-mole task with identical statistical properties to the human participants. We chose this architecture for its ability to effectively learn long timescale dependencies (Shewalkar, Nyavanandi, & Ludwig, 2019). Like humans, the model was probed on its performance on location prediction and judgments of similarity among the games. The model had a 1 × 17 input layer, a 1 × 150 layer of gated recurrent hidden units (Chung, Gulcehre, Cho, & Bengio, 2014), and a 1 × 9 output layer (Figure 5A). Adjacent layers were fully connected. The 1 × 17 input vector consisted of eight context units denoting the one-hot encoded game ID and 9 one-hot encoded location units (representing one center and eight peripheral target locations). Gated recurrent units learn weights to generate a candidate hidden activation based on their input and their own previous activation, and then separately learn how much to integrate the candidate hidden activation with the previous activation (again based on both the input and the previous activation) to generate their output. The model was implemented in Python 3.6.10 using keras tensorflow (tensorflow version 2.1) with default fitting parameters and a mean squared error loss function in batches of eight games at a time (equivalent to a single round of exposure to all eight games). Activations were reset in between games. The effective learning rate was adjusted automatically over training using the RMSprop optimizer to speed convergence. Error was computed by comparing the predicted location encoded in the 1 × 9 output vector against the actual subsequent location.

Training

Models were trained via Backpropagation Through Time on sequences generated using the same randomization and counterbalancing procedures as human participants, for 240 rounds of games. Preliminary examination suggested that the model performed similarly to human participants on prediction and similarity judgment tasks after receiving approximately 5 times the exposure that human participants did (40 rounds of games), given a default global learning_rate parameter of 0.001. Presented analyses probe the model after 5 times human exposure, unless otherwise specified. Separate model instances were initialized with random weights as follows: Connection weights between layers were taken from a Glorot uniform distribution (Glorot & Bengio, 2010), recurrent weights were computed from an orthogonal matrix derived from an initial random normal matrix, and bias weights were initially set to zero.

Analysis

Online predictions. At various stages of training, weights were frozen and the model was asked to predict the next location at the same probe points as the human participants on a number of trials (172,800 trials over 800 rounds of the eight games). Max activation of the output units was used to determine the most likely predicted location. The same sensitivity measures were derived as for humans: low-level sensitivity, and high-level sensitivity at the earliest probe point in each game. Gaussian noise was applied to all 17 input units before making a prediction. Several noise levels were assessed (0.01, 10, and 0.1–2.0 in increments of 0.1), and a Gaussian with a standard deviation of 1.0 was determined to best correspond to human data. To create stable performance estimates, performance was averaged over 10 injections of noise.

In addition to comparing the model's performance to human behavior, we also examined its underlying

predictive mechanisms using additional online prediction tests. We used Layer-wise Relevance Propagation (LRP) to estimate to what extent trials at different lags in the past influenced model predictions (Appendix A). We also assessed whether high-level sensitivity in particular is explainable based on recent past input. To this end, we tested whether the model could apply the high-level rule correctly at the start of the second half of gameplay, and whether our mid-game event boundary marker (four center location trials, rather than one) was a sufficient cue to obtain high accuracy at that point. Thus, we checked high-level accuracy of the trained model on the first noncenter trial of the second half of gameplay. To simulate elimination of preceding context, we also checked high-level accuracy on the first noncenter trial at the start of novel games that began their first half with four center trials, rather than one. We did not apply noise for these assessments, and we ignored output activation for the center location (analogous to participants not having the option of selecting the center location).

Similarity judgments. After training, model weights were frozen and the model was exposed to inputs such that a single game ID unit was turned on and all other input units were off. Activation of the model's hidden units was compared across games using Pearson correlations as a metric of distance. The model was then asked to compare games' similarity using trials generated in the same way as for human judgments. A softmax function was used to determine the probability of each choice based on its relative distance to target (temperature = 1), and we sampled from the probability distribution 10 times for each trial to generate probabilistic choice data and then took the average performance across trials for each trial type and model instance. Trial types were analyzed analogously to the human data.

Activation trajectories. Hidden unit activations were recorded during gameplay. The average hidden activation vector was taken across all rounds for each game ID and trial (from 1–27 during the game) for each model instance. Then distance matrices between all the averaged activity vectors were computed for each model instance, and the distance matrices were averaged across model instances. Multidimensional scaling (in two dimensions) was applied to the final aggregate distance matrix using the sklearn.manifold.MDS function from the scikit-learn package (version 0.24.2) in Python (version 3.6.10) with default parameters, and the resulting mean activation trajectories over the course of gameplay were plotted.

RESULTS

Human Behavior

Online Predictions

Following our preregistered analysis plan, we fit logistic regression models to the online prediction data (Figure 2A). In our assessment of high-level sensitivity, participants were above chance at predicting locations from the correct half of the game (earliest probe point: mean = .266, SE = 0.060, Z ratio = 4.424, $p < .001$; all probe points: mean = .525, SE = .053, Z ratio = 9.891, $p < .001$) and did so more often later in training, although not significantly so when only the earliest probe point was considered (earliest probe point: $\beta = .310$, $SE = .195$, $\chi^2_1 = 2.525$, *ns*; all probe points: $\beta = .410$, $SE = .150$, $\chi^2_1 = 7.492$, $p = .006$). Participants were also more likely to select the appropriate second location in a pair than the location corresponding to the opposite low-level order rule ($\beta = .892$, $SE = .112$, Z ratio = 7.953, $p < .001$). Again, this trend increased over the course of training ($\beta = 1.605$, $SE = .326$, $\chi^2_1 = 24.204$, $p < .001$).

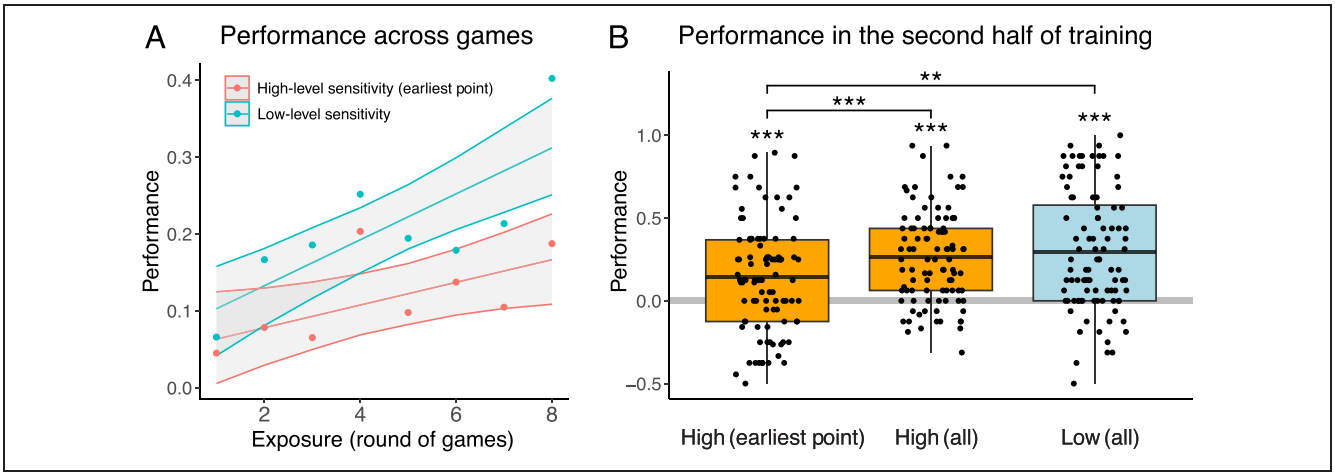

Figure 2. Online performance measures of learning in human participants. (A) Sensitivity to high- and low-level order information plotted by round of eight games over the course of training. Shaded regions indicate bootstrapped 95% confidence bands. (B) Participants demonstrated sensitivity to high- and low-level order rules in their online predictions of upcoming locations during the second half of training. Box plot midlines show the mean, and error bars indicate upper and lower quartiles, $^{**}p < .01$, $^{***}p < .001$.

Aggregating across the second half of training, humans were above chance on our measures of both low- and high-level sensitivity, low-level sensitivity: mean = .298, $t(95) = 8.073$, $p < .001$; high-level sensitivity (all trials): mean = .264, $t(95) = 9.305$, $p < .001$; high-level sensitivity (earliest trial in game): mean = .142, $t = 4.222$, $p < .001$ (Figure 2B). High-level performance was higher measured across all trials than at the earliest trial in game, difference = .122, $t(95) = 5.550$, $p < .001$. The earliest trial in game, or "earliest probe point," refers to the first noncenter location trial, or the second absolute trial in each game. Low-level performance did not differ from high-level performance across all trials (difference = −.034, $t = -.947$, *ns*). However, low-level performance (across all trials) was higher than high-level performance at the earliest trial in game (difference = .155, $t(95) = 3.295$, $p = .001$).

Overall, participants learned to predict locations in a manner indicating sensitivity to both low- and high-level structure, with some evidence for stronger low-level sensitivity.

Similarity Judgments

Instructions for the similarity judgment task were manipulated such that half of the participants were asked to judge similarity based on "when/where" the target appeared, and half were asked to judge based on the "sequence of locations" that the target appeared in. Performance was marginally better when participants were asked to base their judgments on "when/where" the target appeared, rather than the "sequence of locations" (three trial type scheme: $\beta = -0.255$, $SE = .134$, $\chi^2_1 = 3.509$, $p = .061$). Adding an interaction term between instruction condition and 3-way trial type—high, low, or both—did not improve model fit ($\chi^2_2 = 2.081$, *ns*). We specifically predicted based on pilot data that instructions condition would modulate sensitivity to low-level order on the similarity judgment task, and this was not born out ($\chi^2_1 = .127$, *ns*). Thus, we collapsed data across instructions condition for subsequent analyses.

Using logistic regression as per our preregistered analysis plan, but now collapsing across instructions condition, human participants were above chance on trials indexing sensitivity to low-level order and to both low- and high-level order combined (low: $\beta = .207$, $SE = .053$, Z ratio = 3.938, $p < .001$; both: $\beta = .278$, $SE = .068$, Z ratio = 4.096, $p < .001$). They were only marginally above chance on trials indexing sensitivity to high-level order ($\beta = .088$, $SE = .052$, Z ratio = 1.674, $p = .094$). Participants had higher accuracy on trials that probed low-level sensitivity than on trials that probed high-level sensitivity only (high vs. low odds ratio = .888, $SE = 0.0378$, Z ratio = −2.793, adjusted $p = .015$; both vs. high odds ratio = 1.209, $SE = .0732$, Z ratio = 3.142, adjusted $p = .005$; p values adjusted by the Tukey method). Trials that probed low-level sensitivity did not differ in accuracy from trials reflecting both high- and low-level sensitivity (both vs. low odds ratio = 1.074, $SE = .065$, Z ratio = 1.177, *ns*).

Broken down further by trial type (five-way scheme), participants were above chance on all trial types except for one measure of high-level sensitivity ("two different rules vs. same high-level rule"): (b) both same rules versus both different rules, estimated mean probability correct = .569, $SE = .0166$, Z ratio = 4.094, $p < .001$; (c) same rules versus low level different, estimated mean probability correct = .563, $SE = .0167$, Z ratio = 3.718, $p < .001$; (d) same rules versus high level different, estimated mean probability correct = .535, $SE = .0168$, Z ratio = 2.074, $p = .038$; (e) both different rules versus only high level differs, estimated mean probability correct = .546, $SE = .014$, Z ratio = 3.250, $p = .001$; (f) both different rules versus only low level differs, estimated mean probability correct = .515, $SE = .0141$, Z ratio = 1.088, *ns*.

Participants showed high accuracy on (a) the attention check, mean = .970, $t(95) = 77.125$, $p < .001$, and were significantly biased toward using low-level order information when high- and low-level information sources were directly pitted against each other (Trial Type g; mean = .458, $t = -2.543$, $p = .013$). The t tests corresponding with Figure 3 are reported below for each trial type: (b) Same rules (but different game identity / target image) versus game with different high- and low-level order condition, mean = .564, $t(95) = 3.451$, $p < .001$. (c) Same rules versus game with different low-level order condition, mean = .558, $t(95) = 3.268$, $p = .002$. (d) Same rules versus game with different high-level order condition, mean = .532, $t(95) = 1.804$, $p = .074$. (e) Both low- and high-level order differ versus only high-level order differs from the comparison game, mean = .543, $t(95) = 2.678$, $p = .009$. (f) Both low- and high-level order differ versus only low-level order differs from the comparison game (i.e., high-level order matches in one condition). Above-chance accuracy would reflect sensitivity to high-level order information, mean = .513, $t(95) = 0.978$, *ns*.

Overall, both high- and low-level order influenced participants' similarity judgments, but judgments were more strongly impacted by low-level order information.

Correlation Analysis

In an exploratory analysis, we further examined by-subject correlations among sensitivity measures from our online prediction and similarity judgment tasks (Figure 4). We tested all possible correlations between all of the above reported behavioral measures across both tasks, and corrected for the false discovery rate (FDR) using Benjamini and Hochberg's (1995) procedure ($q = 0.05$; reported p values and confidence intervals are uncorrected, but all significant results survived FDR correction). We found that measures of sensitivity to low-level order information tended to be correlated with each other across tasks, and the same was true for high-level order information.

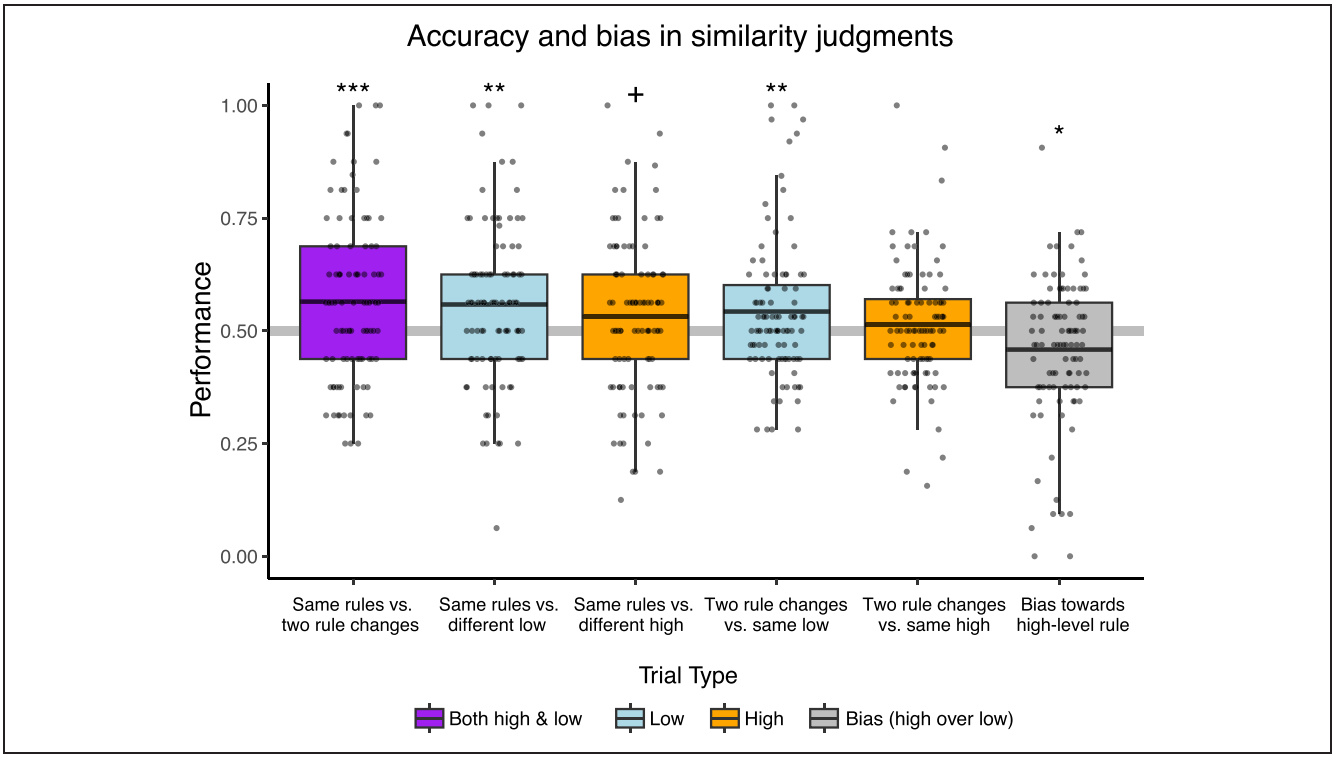

Figure 3. Human similarity judgments. Participants were more sensitive to low- than high-level order information in their similarity judgments. Box plot midlines show the means, $+p < .1$, $*p < .05$, $**p < .01$, $***p < .001$.

Specifically, sensitivity to low-level order on the online prediction task was correlated with both measures of low-level sensitivity on the similarity judgment task ("same rules vs. different low": $r = .321$, 95% CI [.128, .490], $t = 3.281$, $p = .001$; "two rule changes vs. same low": $r = .319$, 95% CI [.127, .488], $t = 3.264$, $p = .002$). Sensitivity to high-level order on early trials in gameplay on the online prediction task was correlated with high-level sensitivity

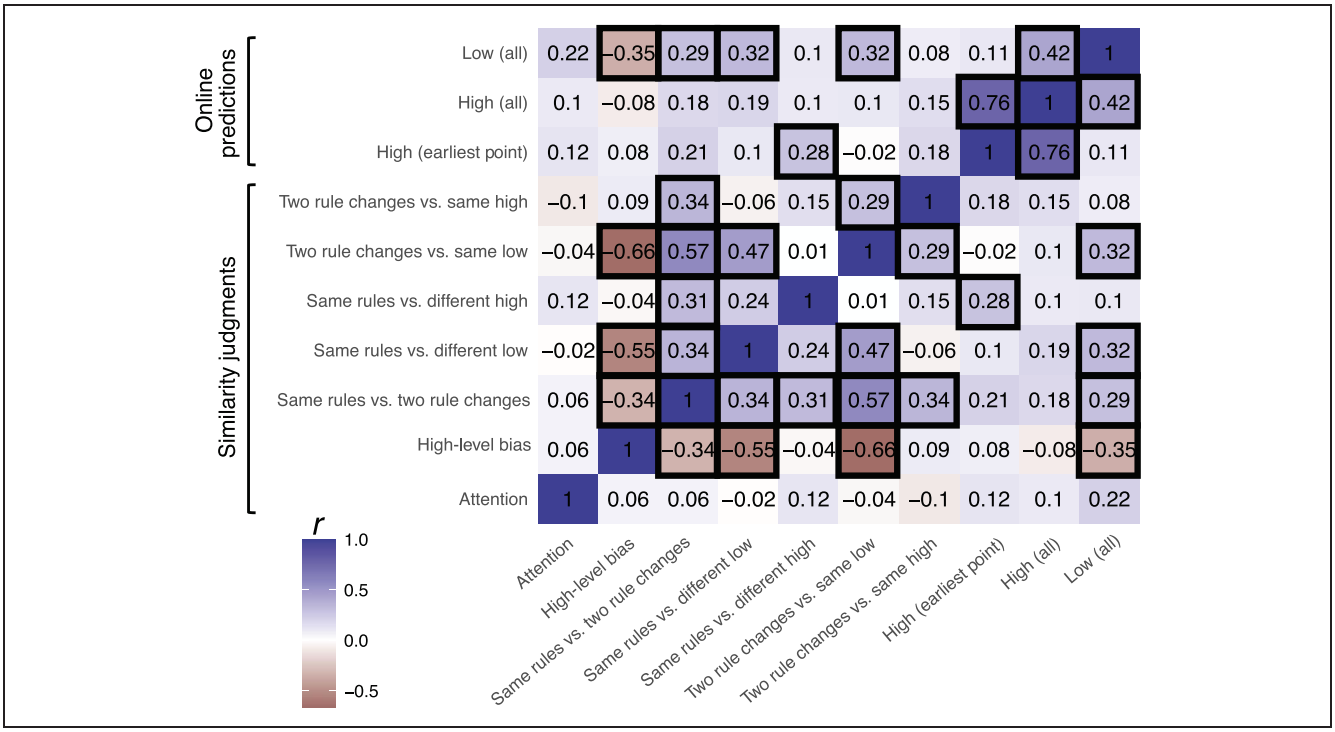

Figure 4. Correlations among human behavioral measures. Pearson correlations surviving FDR correction ($q < .05$) outlined in black.

on "same rules vs. different high" trials in the similarity judgment task (r = .280, 95% CI [.084, .455], t = 2.824, p = .006). Although there was a significant positive correlation between one measure of online sensitivity to high-level order and online sensitivity to low-level order (r = .422, 95% CI [.242, .574], t = 4.517, p < .001), the purer measure of high-level sensitivity (at the earliest points in gameplay) did not correlate with low-level order information (the measures in Figure 2B; r = .106, 95% CI [−.096, .300], Bayes factor against r = .333 was .388), suggesting some degree of decoupling of high- and low-level sensitivity.

Neural Network Model

Online Predictions

At 40 epochs of training (exposure to 320 games), the model showed above-chance low-level sensitivity, mean = .400, $t(47)$ = 31.145, p < .001, and high-level sensitivity, mean = .239, $t(47)$ = 85.008, p < .001, in its predictions. Like human participants, the model demonstrated higher performance on our measure of low-level sensitivity than on high-level sensitivity over the course of learning (Figure 5B), difference by 40 rounds exposure = .161, $t(47)$ = 11.479, p < .001. To match the human data, high-level sensitivity was computed at the earliest noncenter trial in each game, but low-level sensitivity was computed throughout the game (for the second location among paired locations).

We also tested model high-level predictions at the start of the second half of gameplay, with and without having the model state informed by the first half of gameplay. After 40 epochs of training, the model demonstrated high-level sensitivity on the first noncenter trial of the second half of gameplay, mean = .995, $t(47)$ = 333.640, p < .001. When the model was presented with four center trials in the absence of any previous input, this sensitivity was eliminated, mean = −.052, $t(47)$ = −1.273, *ns*, although it quickly adjusted to local context and always made perfect high-level predictions by the second noncenter location pair.

Similarity Judgments

The model also showed a pattern of similarity judgments that was qualitatively similar to human participants (Figure 5C). Individual tests of above-chance accuracy

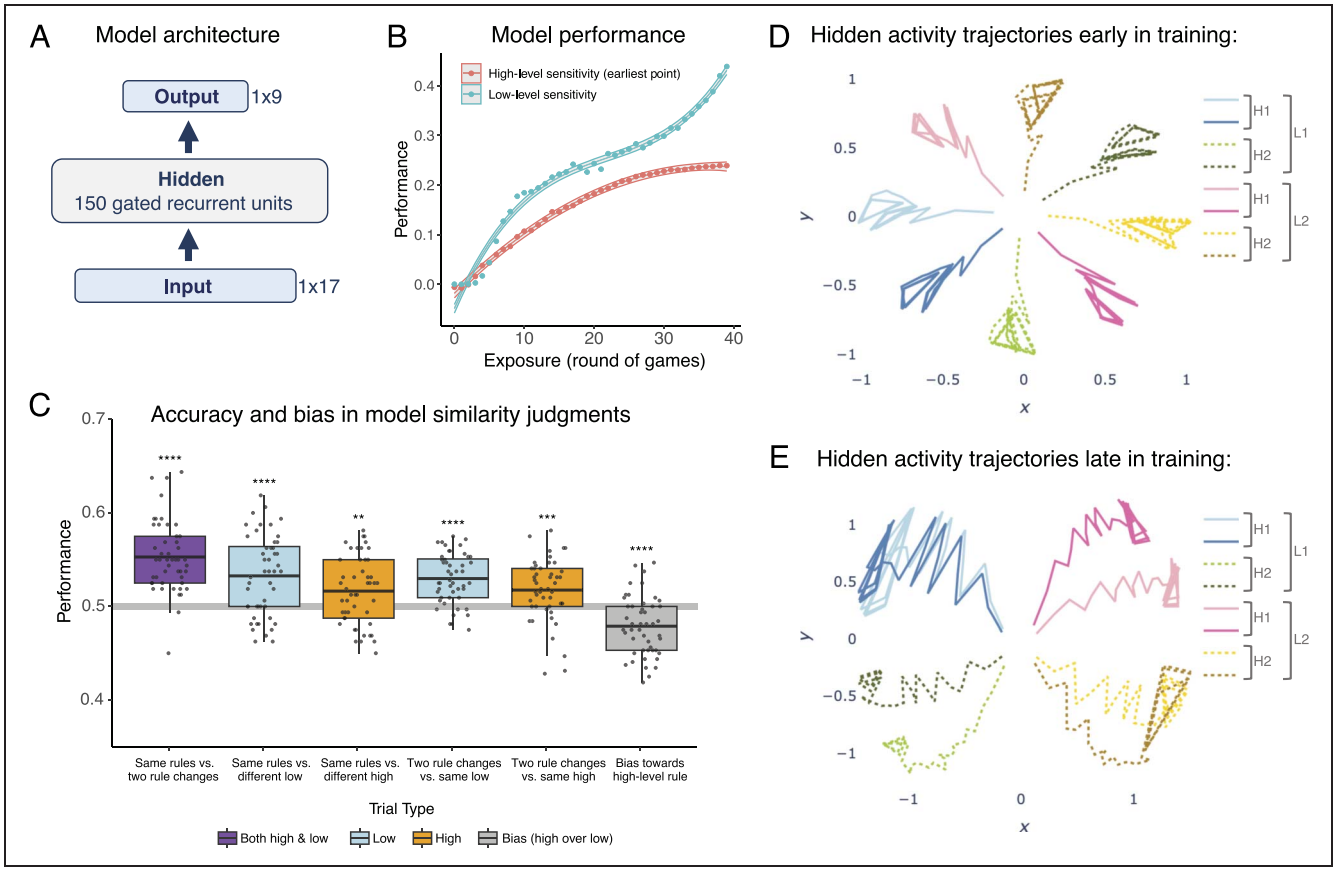

Figure 5. Modeling results. (A) Model architecture. (B) Average model performance on online prediction task over training. Shaded regions indicate bootstrapped 95% confidence bands. (C) Model similarity judgments after 40 rounds of training. Box plot midlines show the means, $^{**}p$ < .01, $^{***}p$ < .001, $^{****}p$ < .0001. (D and E) Mean hidden activation trajectories during the course of gameplay for each of the eight games early in training (D, eight rounds), and late in training (E, 40 rounds). Hidden activation trajectories start out roughly equidistant among the eight games but become clustered by high- and low-level order rules over the course of training.

and high- versus low-level bias are reported below for each trial type: (b) Same rules (but different game identity / target image) versus game with different high- and low-level order condition, mean = .553, $t(47) = 9.682$, $p < .001$. (c) Same rules versus game with different low-level order condition, mean = .533, $t(47) = 5.273$, $p < .001$. (d) Same rules versus game with different high-level order condition, mean = .516, $t(47) = 3.143$, $p = .003$. (e) Both low- and high-level order differ versus only high-level order differs from the comparison game. Above-chance accuracy would reflect sensitivity to low-level order information, mean = .530, $t(47) = 7.759$, $p < .001$. (f) Both low- and high-level order differ versus only low-level order differs from the comparison game. Above-chance accuracy would reflect sensitivity to high-level order information, mean = .518, $t(47) = 3.661$, $p = .001$. (g) High-level bias. One game matches the comparison game on low-level order, and the other matches the comparison game on high-level order, mean = .479, $t(47) = -4.645$, $p < .001$.

Accuracy was higher on trials probing low-level than high-level sensitivity (mean difference = .014, $t(47) = 2.767$, $p = .008$).

Model Hidden Activation Trajectories

We examined the model's hidden activation trajectories during gameplay for each game ID at both early (eight rounds of games; Figure 5D) and late (40 rounds; Figure 5E) stages of training. Trajectories are shown branching out from the center of the plot as gameplay continues from the start to end of each game. Early in training, the model represented individual games differently, and all games were approximately equally dissimilar (relative to changes over time within the game). Later in training, the model grouped games by high- and low-level order rule. Games that differed in their low-level order rule were more distinct in activation space than games that differed in their high-level order rule (mean of [[distance for different low level rule – same low level rule] – [distance for different high level rule – same high level rule]] = .396, $t(47) = 6.103$, $p < .001$). Stronger representational warping according to low-level order corresponds to the stronger sensitivity to low-level order observed in the model's similarity judgments as well as online predictions.

Model activation trajectories changed gradually over the course of gameplay, consistent with sensitivity to the distant past. To confirm this interpretation, we applied LRP and found that model predictions were sensitive to input at both near and far (>4) temporal lags (Appendix A).

DISCUSSION

In this study, we have shown that humans can rapidly learn statistical information at slow (high-level) temporal scales while performing a task that focuses on concurrent fast (low-level) statistics. We have also captured aspects of human behavior using a recurrent neural network trained to predict immediate upcoming input, which recapitulated patterns in human online predictions and similarity judgments. The modeling results suggest that a common learning mechanism and representational substrate can capture information at multiple temporal scales. Both the human and modeling results were consistent with prioritization of learning for rapid timescale statistical relationships over slower timescale relationships and with learning at the two levels unfolding in parallel.

Humans were able to learn context-dependent high-level order information while performing a visuo-motor task (whack-a-mole) in which they also learned low-level order information. Notably, because rule-agnostic center location trials were distributed throughout the game, participants could not rely solely on the current visual stimulus to correctly predict upcoming locations. Participants' low-level order knowledge was demonstrated both in their online predictions during training and in their posttraining similarity judgments. Their high-level order knowledge was demonstrated in their online predictions (even when based on the visual context alone) and on similarity judgment trials that pitted a match on high- and low-level order against a high-level mismatch. It was also reflected in the positive correlation between high-level order sensitivity on the online prediction task (on early trials using visual context) and the similarity judgment task. In both online predictions and similarity judgments, there was evidence for greater sensitivity to low-level statistics.

The fact that humans were able to rapidly gain sensitivity to high-level order statistics while performing a task focused on learning low-level order statistics is nontrivial, because the presence of low-level dependencies has been suggested to interfere with learning of higher-level dependencies in some domains (e.g., Newport & Aslin, 2004; Gómez, 2002). Our findings are consistent with faster temporal statistics being represented differently enough from these kinds of slower temporal statistics during visuomotor learning that they do not provoke interference. It may be that high-level order information unfolding at longer timescales and with more intervening trials (relative to the long-distance dependencies employed in, e.g., Gómez, 2002, which involved separation by a single word) allows more independence of learning across timescales. There are other differences between the paradigms that could be relevant, however, such as our high-level states having looser order constraints on the trials nested within them (i.e., four pairs of locations were presented in each half of the game, and the order of the four pairs was randomized each time a game was played). Gated recurrent networks similar to our model should be able to handle both the shorter and longer variants of long-distance dependencies (Chung et al., 2014), and it would be interesting for future simulation work to explore under what timescales or other conditions interference may arise. A degree of dissociability in the representations of slow and fast temporal statistics is not incompatible with the

idea that both could be acquired using a common learning mechanism and brain region, as suggested by our modeling results. An individual region can learn to rely on somewhat distinct populations of neurons to process different timescales of information.

A gated recurrent neural network with a single hidden layer was able to learn statistical patterns unfolding at slow and fast timescales, displaying sensitivity to both the recent and more distant past, and approximated human behavior in our paradigm. The model made online predictions that matched human low- and high-level sensitivity measures, with an advantage for low-level predictions, and, like humans, demonstrated greater sensitivity to low- than high-level order information in the similarity judgment task. The close match to the human data at multiple timescales is notable given that our model was only trained to predict the upcoming target location. Training our model to predict the upcoming target location may have resulted in a close match to our human data because it matches our explicit predictive probe task, and/or because it is consistent with a broader prediction-based learning account of statistical learning (e.g., Schapiro et al., 2013; Kiebel, Daunizeau, & Friston, 2008; Friston, 2005). It is also interesting that our model was able to learn statistical dependencies at multiple timescales given that it uses a single, unified learning mechanism. Our model learns what temporal scales are relevant, rather than having them explicitly parameterized, and accomplishes learning with a single hidden layer. This is in contrast with models such as the Hierarchical Autoencoders in Time model (Chien & Honey, 2020), which employs a more constrained gating mechanism that is modulated by layer depth and explicit temporal integration parameters to learn statistics at multiple temporal scales. Our model's success despite its flat structure is consistent with prior demonstrations that single layer models can capture complex multilevel temporal dynamics (e.g., Botvinick, 2007; Botvinick & Plaut, 2004).

Our model also demonstrates that slow and fast timescale statistical relationships can be learned concurrently. Although the model did show bias toward enhanced learning of low-level / fast temporal statistics, it improved on sensitivity measures for both fast and slow timescale statistics simultaneously. This matches visual trends in the human data (Figure 2B), although we were not powered to examine this directly. Simultaneous learning at multiple timescales is in contrast with suggestions that fast temporal relationships must be learned before slower ones in the motor (Krakauer et al., 2019) and language (Saffran & Wilson, 2003) learning literatures. It is possible that previous findings may reflect a bias in the strength of (simultaneous) learning for statistics that span different temporal scales, rather than a system constrained to learn rules at different timescales in a strictly sequential fashion. The lack of a strong correlation between human low- and high-level sensitivity in the online prediction task is also consistent with this. Future work will be needed to confirm differences in the learning rate for statistical learning of input that spans different timescales, and confirm whether information presented at different timescales embedded in a single motor-perceptual stream is truly acquired simultaneously in humans. In addition, more work will be needed to tease apart the role of timescale per se versus conceptual complexity and level of abstraction in explaining variance in learning trajectories for different temporal scales.

Our modeling results speak to how statistical dependencies at multiple timescales may be learned in the brain. We have previously developed a model of the hippocampus that provides an account of its role in rapid statistical learning (Schapiro, Turk-Browne, Botvinick, & Norman, 2017). In the model, the monosynaptic pathway to region CA1 acts as a neural network with a single hidden layer, employing distributed representations and a relatively fast learning rate that allow it to effectively learn short timescale statistics quickly. Our current modeling results suggest that such a single-hidden-layer system may be able to concurrently handle longer timescale statistics. Temporal dependencies are known to be encoded at multiple timescales in both neocortex (Baldassano et al., 2017; Hasson, Chen, & Honey, 2015; Lerner, Honey, Silbert, & Hasson, 2011) and the hippocampus (reviewed in Davachi & DuBrow, 2015). In both cases, there appears to be an anatomical gradient of sensitivity to different timescales in different areas, with the hippocampus exhibiting increasing sensitivity to long timescales moving more anteriorly/ventrally along its long axis (Bouffard et al., 2023; Tarder-Stoll, Baldassano, & Aly, 2023; Raut, Snyder, & Raichle, 2020; Brunec et al., 2018). It may be that recurrent machinery allowing sensitivity to longer timescale statistics is increasingly present in more anterior/ventral segments of the hippocampus and/or its inputs. Neocortical gradients of timescale dependency may emerge for longer term forms of temporal knowledge.

In conclusion, humans are able to learn statistical information at multiple timescales within a short period, and their behavior can be effectively modeled using recurrent neural networks. Rather than learning rapid timescale statistics as a prerequisite for learning slower statistics, our model suggests that statistical information can be learned across multiple timescales simultaneously and via a shared mechanism and substrate. Although we may acquire rapid temporal regularities more readily than slowly evolving ones, the work demonstrates that learning one is not always a prerequisite for or barrier against learning the other.

APPENDIX A: LAYERWISE RELEVANCE PROPAGATION

To probe how far into the past input influences model predictions for the upcoming location, we used an explainable AI decomposition method called Layerwise Relevance Propagation (LRP) (Bach et al., 2015). In LRP, relevance scores are computed and assigned to hidden units and then to upstream input units, on a trial-by-trial

basis. Scores are computed for simple units in linear proportion to magnitude of excitatory and inhibitory inputs and their weights. Specifically:

$$R_{i\leftarrow j}^{(l,l+1)} = \begin{cases} \dfrac{Z_{ij}}{Z_j + \varepsilon} \cdot R_j^{(l+1)} & Z_j \geq 0 \\ \dfrac{Z_{ij}}{Z_j - \varepsilon} \cdot R_j^{(l+1)} & Z_j < 0 \end{cases} \quad (1)$$

where $R_{i\leftarrow j}^{(l,l+1)}$ refers to the relevance of input unit i in layer l to the output activation of unit j in layer $l + 1$, z_{ij} is the linear activation of unit i times the connection weight from unit i to j, z_j is the linear activation of unit j, ε is a small positive stabilization constant, and $R_j^{(l,l+1)}$ is the relevance assigned to unit j in layer $l + 1$. In the case of recurrent neural networks, scores are assigned in a similar manner but account for the multiple input matrices and weight transformations internal to each unit. Relevance was not redistributed to bias parameters for the purposes of this analysis.

LRP has some limitations; for example, relevance heatmaps contain noise and nondiscriminative information when applied to convolutional neural networks trained on image classification (Jung, Han, & Choi, 2021). However, its simple formulation allows it to be applied to a variety of network architectures. In the case of recurrent neural networks, relevance propagation can proceed not only across layers, but also backward through time, shedding light on the relative contributions of points in the past to the current prediction. For this reason, we have assessed the "relevance" assigned to points at various lags in the past as an approximation of the model's forgetting curve. Because relevance scores can be either positive (suggesting excitation) or negative (suggesting inhibition), we took the sum over input units of the absolute values of the relevance scores (broken down into "location" and "game ID" input groups) for all lags during gameplay on a large test set (6400 games; Figure A1). Ninety-five percent bootstrapped confidence intervals (n = 3000 simulations) were computed across 48 model instances.

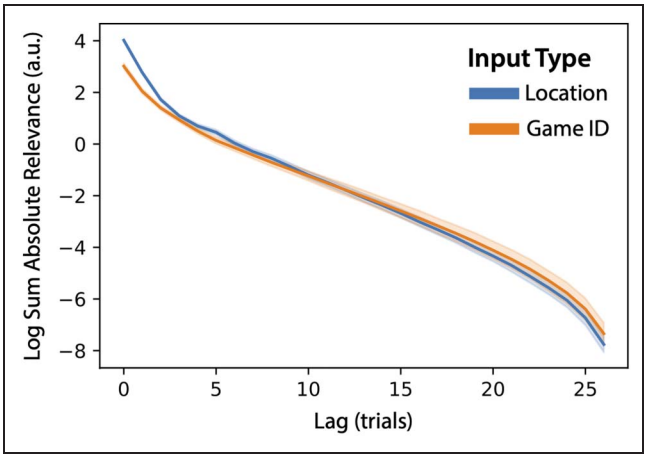

Figure A1. Log sum of absolute relevance scores assigned to location and game ID units over time. Lag 0 signifies input on the current trial. Error bars indicate 95% confidence intervals.

The model assigns relevance preferentially to points in the recent past. However, some relevance is assigned to both location and game ID inputs going back many trials into the past.

Acknowledgments

The authors thank Joey Zhao for assistance with data collection and Elisabeth Karuza for helpful discussions.

Corresponding author: Cybelle M. Smith, Department of Psychology, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104-6243, or via e-mail: cybelle@sas.upenn.edu.

Data Availability Statement

Data and analysis scripts are available at Harvard Dataverse: https://doi.org/10.7910/DVN/GULYJP.

Author Contributions

Cybelle M. Smith: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. Sharon L. Thompson-Schill: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Writing—Review & editing. Anna C. Schapiro: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Writing—Review & editing.

Funding Information

This work was supported by National Institutes of Health (https://dx.doi.org/10.13039/100000002), grant number: F32MH123002 to C. M. S., R01 DC00920 to S. L. T.-S., and R01 MH129436 to A. C. S.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when

selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by Layer-wise Relevance Propagation. *PLoS One*, *10*, e0130140. https://doi.org/10.1371/journal.pone.0130140, PubMed: 26161953

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*, 709–721. https://doi.org/10.1016/j.neuron.2017.06.041, PubMed: 28772125

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: A model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *362*, 1615–1626. https://doi.org/10.1098/rstb.2007.2056, PubMed: 17428777

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395–429. https://doi.org/10.1037/0033-295X.111.2.395, PubMed: 15065915

Bouffard, N. R., Golestani, A., Brunec, I. K., Bellana, B., Park, J. Y., Barense, M. D., et al. (2023). Single voxel autocorrelation uncovers gradients of temporal dynamics in the hippocampus and entorhinal cortex during rest and navigation. *Cerebral Cortex*, *33*, 3265–3283. https://doi.org/10.1093/cercor/bhac480, PubMed: 36573396

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z.-X., et al. (2018). Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Current Biology*, *28*, 2129–2135. https://doi.org/10.1016/j.cub.2018.05.016, PubMed: 29937352

Chien, H.-Y. S., & Honey, C. J. (2020). Constructing and forgetting temporal context in the human cerebral cortex. *Neuron*, *106*, 675–686. https://doi.org/10.1016/j.neuron.2020.02.013, PubMed: 32164874

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*. https://doi.org/10.48550/arXiv.1412.3555

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253. https://doi.org/10.1037/0096-3445.120.3.235, PubMed: 1836490

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1119–1130. https://doi.org/10.1037/0278-7393.30.5.1119, PubMed: 15355140

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, *19*, 92–99. https://doi.org/10.1016/j.tics.2014.12.004, PubMed: 25600586

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458–467. https://doi.org/10.1037/0278-7393.28.3.458, PubMed: 12018498

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *360*, 815–836. https://doi.org/10.1098/rstb.2005.1622, PubMed: 15937014

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., & Griffiths, T. D. (2011). Neural prediction of higher-order auditory sequence statistics. *Neuroimage*, *54*, 2267–2277. https://doi.org/10.1016/j.neuroimage.2010.10.038, PubMed: 20970510

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256). https://proceedings.mlr.press/v9/glorot10a

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436. https://doi.org/10.1111/1467-9280.00476, PubMed: 12219809

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*, 304–313. https://doi.org/10.1016/j.tics.2015.04.006, PubMed: 25980649

Jung, Y.-J., Han, S.-H., & Choi, H.-J. (2021). Explaining CNN and RNN using selective layer-wise relevance propagation. *IEEE Access*, *9*, 18670–18681. https://doi.org/10.1109/ACCESS.2021.3051171

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*, 12733. https://doi.org/10.1038/s41598-017-12876-5, PubMed: 28986524

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, *4*, e1000209. https://doi.org/10.1371/journal.pcbi.1000209, PubMed: 19008936

Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., & Haith, A. M. (2019). Motor learning. *Comprehensive Physiology*, *9*, 613–663. https://doi.org/10.1002/cphy.c170043, PubMed: 30873583

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *eLife*, *10*, e64972. https://doi.org/10.7554/eLife.64972, PubMed: 33884953

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011, PubMed: 21414912

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 523–530. https://doi.org/10.1037/0278-7393.13.4.523

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *1*, 31. https://doi.org/10.3389/fpsyg.2010.00031, PubMed: 21833201

Momennejad, I. (2024). Memory, space, and planning: Multiscale predictive representations. *arXiv*. https://doi.org/10.48550/arXiv.2401.09491

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162. https://doi.org/10.1016/S0010-0285(03)00128-2, PubMed: 14732409

Raut, R. V., Snyder, A. Z., & Raichle, M. E. (2020). Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proceedings of the National Academy of Sciences, U.S.A.*, *117*, 20890–20897. https://doi.org/10.1073/pnas.2003383117, PubMed: 32817467

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926, PubMed: 8943209

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*, 273–284. https://doi.org/10.1207/S15327078IN0402_07

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*, 229–242. https://doi.org/10.1007/s00221-003-1548-8, PubMed: 12879170

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486–492. https://doi.org/10.1038/nn.3331, PubMed: 23416451

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *372*, 20160049. https://doi.org/10.1098/rstb.2016.0049, PubMed: 27872368

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*, 235–245. https://doi.org/10.2478/jaiscr-2019-0006

Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science*, *13*, 106–127. https://doi.org/10.1111/tops.12505, PubMed: 32459391

Tarder-Stoll, H., Baldassano, C., & Aly, M. (2023). The brain hierarchically represents the past and future during multistep anticipation. *bioRxiv*. https://doi.org/10.1101/2023.07.24.550399, PubMed: 37546761