

---

# Interval timing in deep reinforcement learning agents

---

**Ben Deverett**  
DeepMind  
bendevert@google.com

**Ryan Faulkner**  
DeepMind  
rfaulk@google.com

**Meire Fortunato**  
DeepMind  
meirefortunato@google.com

**Greg Wayne**  
DeepMind  
gregwayne@google.com

**Joel Z. Leibo**  
DeepMind  
jzl@google.com

## Abstract

The measurement of time is central to intelligent behavior. We know that both animals and artificial agents can successfully use temporal dependencies to select actions. In artificial agents, little work has directly addressed (1) which architectural components are necessary for successful development of this ability, (2) how this timing ability comes to be represented in the units and actions of the agent, and (3) whether the resulting behavior of the system converges on solutions similar to those of biology. Here we studied interval timing abilities in deep reinforcement learning agents trained end-to-end on an interval reproduction paradigm inspired by experimental literature on mechanisms of timing. We characterize the strategies developed by recurrent and feedforward agents, which both succeed at temporal reproduction using distinct mechanisms, some of which bear specific and intriguing similarities to biological systems. These findings advance our understanding of how agents come to represent time, and they highlight the value of experimentally inspired approaches to characterizing agent abilities.

## 1 Introduction

To exploit the rewards available in our environment, we capitalize on relationships between environmental causes and effects that exhibit precise temporal dependencies. For example, to avoid a dangerous threat moving towards you, you may estimate its speed by observing its displacement over a fixed time interval, extrapolate its future position over another time interval, and condition your escape behavior on your estimated time of contact with the threat. This ability to measure time and use it to guide behavior is necessary and prevalent in both animals and artificial agents. However, owing to basic differences in their implementation, artificial intelligence (AI) and biology have different relationships to time. Nevertheless, it is likely that consideration of the temporal measurement problem across these domains may yield valuable insights for both.

In biological systems, time measurements are necessary at a variety of temporal scales, ranging from milliseconds to years. The mechanisms underlying these timing abilities differ according to time scale, and many are well characterized at the level of the neural circuits (7; 22). On the scale of seconds, interval timing paradigms are used in animals to study the behavioral and neural properties of time measurement. For example, an animal might be taught to measure out the elapsed interval between two events, then to report or reproduce that interval to the best of their ability in order to obtain a reward (2). Humans, non-human primates, and rodents exhibit a number of characteristic behaviors on these tasks (4; 13; 12) that may reflect biological constraints on mechanisms that remain incompletely understood.

In the AI domain, there exist numerous agents that have succeeded in solving tasks with complex temporal dependencies (29; 27; 28; 10; 11). Many of these are in the category of deep reinforcement learning agents, which develop reinforcement learning policies that use deep neural networks as function approximators (20). While the abilities of these agents have advanced dramatically in recent years, we lack detailed understanding of the solutions they employ.

For instance, consider an agent that must learn to condition its actions on the amount of elapsed time between two environmental stimuli, as we will do in this study. A deep reinforcement learning agent with a recurrent module (e.g. LSTM (9)) has, by construction, two distinct mechanisms for storing relevant timing information. First, the LSTM is a source of temporal memory, since it is designed to store past information in model parameters trained by way of backpropagation through time. Second, the reinforcement learning algorithm, regardless of the underlying function approximator, assigns the credit associated with rewards to specific past states and actions. A deep reinforcement learning agent without a recurrent module (i.e. purely feedforward) lacks the former mechanism but retains the latter one. When trained end-to-end on a timing task, it is unclear whether and how agents may come to implicitly or explicitly represent time.

Here we use an experimentally inspired approach to study the solutions that reinforcement learning agents develop for interval timing. We characterize how the strategies developed by the agents differ from one another, and from animals, discovering themes that underlie interval timing. We suggest that this approach offers benefits both for AI – by introducing an experimental paradigm that simply and precisely evaluates agent strategies, and for neuroscience – by exploring the space of solutions that develop outside of biological constraints, serving as a testbed for interpretation of timing-related findings in animals.

## 2 Methods

### 2.1 Interval reproduction task

We designed a task based on a temporal reproduction behavioral paradigm in the neuroscience literature (13). The task was implemented in PsychLab (17), a simulated laboratory-like environment inside DeepMind lab (1) in which agents view a screen and make “eye” movements to obtain rewards. We have open-sourced the task (along with other related timing tasks) for use in future work.

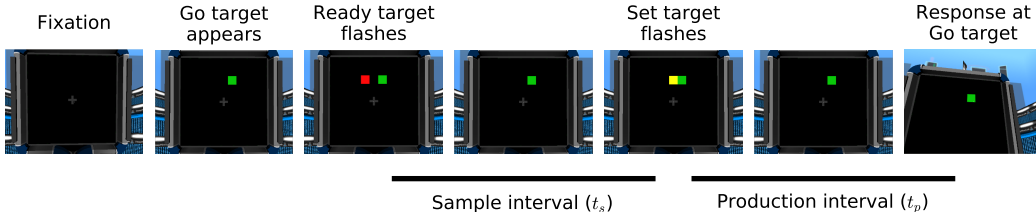


Figure 1: Interval reproduction task. The image sequence shows a single trial of the task. First, the agent fixates on a center cross, at which point the “Go” target appears. After a short delay, the red “Ready” cue flashes, followed by a randomly chosen “sample interval” delay. After the sample interval passes, the yellow “Set” cue flashes. Then the agent must wait for duration of the sample interval before gazing onto the “Go” target to end the trial. If the period over which it waited, the “production interval”, matches the sample interval within a tolerance, the agent is rewarded. This task is closely based on an existing temporal reproduction task for humans and non-human primates (13).

The task is shown in Fig. 1. In each trial, the agent fixates on a central start position, at which point a “Go” target appears on the screen, which will serve as the eventual gaze destination to end the trial. After a delay, a “Ready” cue flashes, followed by a specific “sample” interval of time, then a “Set” cue flashes. Following the flash of the “Set” cue, the agent must wait for the duration of the sample interval before gazing onto the “Go” target to complete the trial. If the duration of the “production” interval (i.e. the elapsed time from “Set” cue appearance until gaze arrives on “Go” target) matches the sample interval within a specified tolerance, the agent is rewarded.

The demands of this “temporal reproduction” task are twofold: the agent must first measure the temporal interval presented between two transient environmental events, and it then must reproduce that interval again before ending the trial. Trials are presented in episodes, with each episode containing 300 seconds or a maximum of 50 trials, whichever comes first. Each trial’s sample interval is selected randomly from the uniform range from 10-100 frames in steps of 10 (corresponding to 167-1667 ms at 60 frames per second). The agent is rewarded if the production interval is sufficiently close to the sample interval; specifically, if  $|t_p - t_s| < \gamma_s(\alpha + \beta t_s)$ , where  $t_p$  is the production interval,  $t_s$  is the sample interval,  $\alpha$  is a baseline tolerance,  $\beta$  is a scaling factor like that used in (13) to account for scalar variability, and  $\gamma_s$  is an overall difficulty scaling factor for each sample interval  $s$ . In practice, we usually set  $\beta$  to 8 frames,  $\alpha$  to zero, and  $\gamma_s$  evolved within an episode from 2.5 to 1.5 to 0, advancing each time two rewards were obtained at the given sample interval  $s$ . In practice we found that the results shown are robust to a wide range of parameters for this curriculum.

## 2.2 Agent architecture

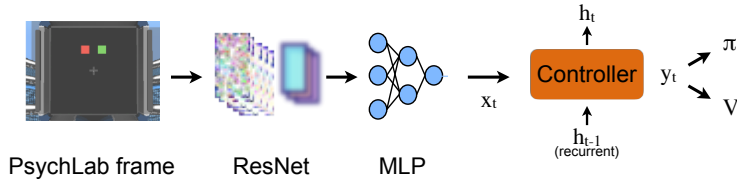


Figure 2: The agent architecture for our interval timing tasks. Frame input is passed into the residual network + MLP. The output from this component is passed to the penultimate component, the controller, which enables the integration of past events in the recurrent case. Finally, this output is sent to the policy and value networks to generate an action (and a policy gradient in the backward pass).

We used an agent based on the A3C architecture (19) (Fig. 2). This agent uses a deep residual network (8) to generate a latent representation of the visual input from Psychlab which is subsequently passed as input to a controller network: either a recurrent network, in this case an LSTM, or a feed-forward network. The controller output is then fed forward to the policy and baseline networks that generate policy and value estimates that are then trained under the Importance Weighted Actor-Learner Architecture (5) (see section A.1). At each time step, the policy generates an action, corresponding to a small instantaneous eye movement in a particular direction from its current position. The LSTM controller provides a way for the agent to integrate past events along with its input in order to drive the policy while in the feed-forward case the agent must rely explicitly on the state of the environment to select actions.

We chose a residual embedding network architecture composed of three convolutional blocks with feature map counts of 16, 32, and 32; each block has a convolutional layer with kernel size 3x3 followed max pooling with kernel size 3x3 and stride 2x2, followed by two residual subblocks. The ResNet was followed by a 256-unit MLP. We used controllers with 128 hidden units for all experiments. The learner was given trajectories of 100 frames, with a batch size of 32, and used 200 actors. Other parameters were a discount factor of 0.99, baseline cost of 0.5, and entropy cost of 0.01. The model was optimized using Adam with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 10^{-4}$ , and a learning rate of  $10^{-5}$ .

## 3 Results

### 3.1 Performance of deep reinforcement learning agents

The recurrent agent learned to perform the task with near-perfect accuracy (Fig. 3a, b; top row); in other words, the production interval was matched to the sample interval across all presented sample intervals. From this analysis, however, it remains unclear whether the agent learned a general timing rule, or whether it memorized a specific discrete set of durations. Fig. 3c demonstrates that the agent indeed learned a general rule, successfully interpolating and extrapolating to new sample intervals on which it was not trained (+ signs in Fig. 3c).

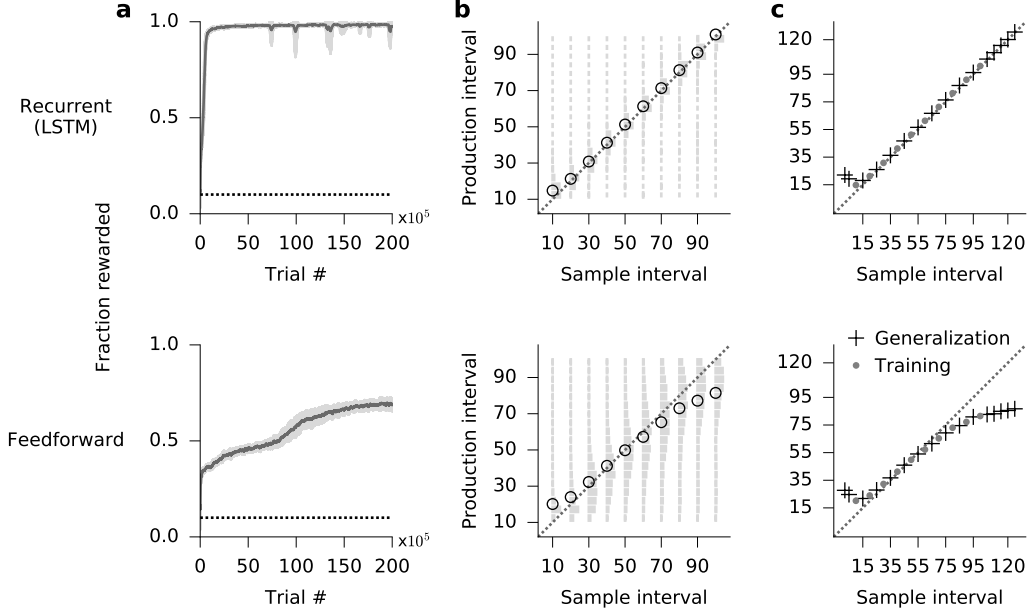


Figure 3: Agent performance on interval reproduction task. (a) Reward rate over training for the recurrent and feedforward agents. Lines show mean  $\pm$  s.d. over 15 seeds. (b) Mean production interval in the trained agent for each of the ten unique sample intervals. Underlying gray histograms show the distribution of production intervals. Includes data from one actor in the final 60,000 trials only, thus excluding the initial training phase. (c) Generalization was assessed by presenting sample intervals not used to train the agent (+ signs).

Somewhat surprisingly, the feedforward agent also learned to performed the task, albeit more slowly and to a lesser degree of accuracy (Fig. 3, bottom row). The feedforward agent exhibited some notable behavioral features relative to the recurrent agent: (1) production interval distributions were wider, (2) a mean-directed bias was found on the production intervals at the extremes of the sample interval distribution, and (3) generalization to untrained intervals was poorer. Nevertheless, the agent learned to produce intervals that were remarkably well matched to the sample intervals, especially given the absence of any traditional or explicit memory systems within the agent architecture.

### 3.2 Psychophysical model of feedforward agent

One notable feature of the feedforward agent’s solution was its similarity to human and primate data (13; 12). In particular, the sigmoid-like shape of the performance curve suggests that the strategy might be well explained by established models of perceptual timing in humans. We tested this intuition quantitatively, since an alignment between agent and animal performance could draw useful links between analyses of agent and animal behaviors.

We fit the feedforward agent data to a Bayesian psychophysical model previously established for human (12) and non-human primate (13) studies. In brief, the model treats the task as a three-stage process: a noisy observation of a sample interval  $t_s$  measured as  $t_m$ , a Bayesian least squares estimation  $t_e$  of the true  $t_s$  given the noisy measurement  $t_m$ , then the generation of a noisy production interval  $t_p$  from the estimated interval  $t_e$ . The measurement and production steps are modeled as Gaussians with one parameter each,  $w_m$  and  $w_p$  respectively, corresponding to the coefficient of variation controlling the scalar variability (6) in the noisy measurement and production processes. The conditional probability of a given production interval  $p(t_p|t_s, w_m, w_p)$  can then be computed by marginalizing over the intermediate distributions as described in the full model, found in (12). The model was fit using optimization routines in Scipy (14).

Fig. 4a shows that in the feedforward agent, the standard deviation of the production interval scales linearly with the sample interval; this feature is known as scalar variability and is used to motivate the psychophysical model because of its prevalence in biological systems (6). Fig. 4b shows the model

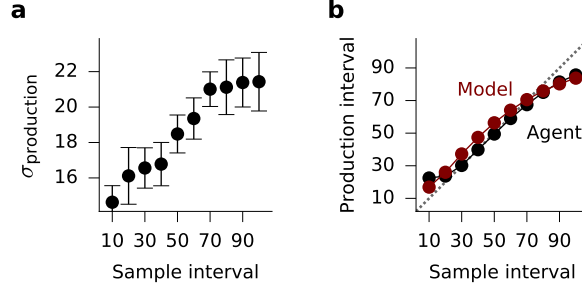


Figure 4: Psychophysical model of interval timing. (a) In the trained agent, the standard deviation of the production interval scales with interval duration. Error bars: s.e.m. over 3 seeds. (b) An established Bayesian psychophysical model was fit to the feedforward agent data.

fit and the agent data; the approximate alignment of these data, and in particular the mean-directed bias at the tails, suggests that this psychophysical model developed for animals may be an appropriate tool for characterizing artificial agent behavior as well.

### 3.3 Evaluation of hidden unit activations

To understand the mechanism used by an agent to solve the task, it is helpful to characterize the activity of the hidden units (29). In this task, it is reasonable to predict that the agent’s hidden units should encode the timing information that the agent has learned in a trial. In particular, in the recurrent agent, one might predict the development of a “timer”, in which the activations of one or many neurons implement a counter that accumulates over the presentation of the sample interval then reads out its value to produce the interval of interest.

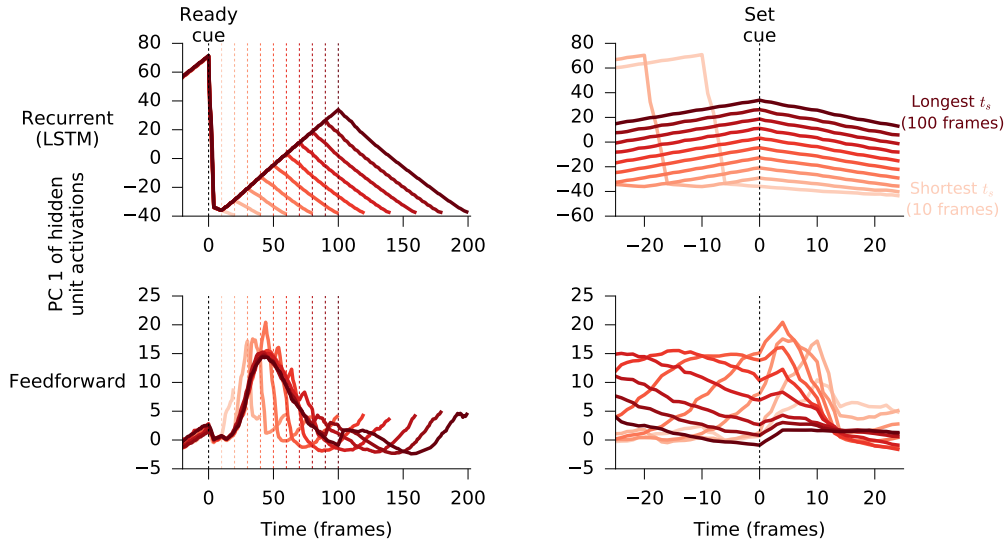


Figure 5: Hidden unit representations of time. The mean activations of the 128 hidden units (top row: LSTM cell state in recurrent agent; bottom row: hidden units in feedforward agent) are shown for trials of each sample interval duration. The activations of the population of hidden units are summarized by the first principal component. Each color corresponds to trials of one specific sample interval duration (darker colors correspond to longer sample intervals). Left column: activations temporally aligned to the Ready cue; colored dashed lines: onset of each respective Set cue. Right column: same data aligned to Set cue.

In Fig. 5 (top row) we show that indeed such counters can be found in the unit activity of the trained recurrent agent. We summarize the unit activity of the 128 LSTM cell state units using their first principal component. During presentation of the sample interval, unit activity rises uniformly, until

the *Set* cue is presented, at which point activity begins to fall, reaching its initial value by the time that duration passes again. This is a simple solution to encoding the time interval and represents a form of clock. As a consequence, it can be seen that the Set-cue-aligned activity separates trials of different duration (Fig. 5, upper right), such that activity falls from a higher set point when the target interval is longer. These aligned average traces bear close resemblance to the unit activity found empirically in non-human primate parietal cortex during interval timing (13).

In feedforward agents, however, it is less clear how the activations of the neurons might represent the timing information. Using the same analysis on the hidden units of the feedforward agent, we found that unit activity also represents intervals, however in a less straightforward way (Fig. 5, bottom). To gain a better understanding of the feedforward solution, we proceeded to analyze the actions of the agent.

### 3.4 Action trajectories

The success of the feedforward agent is intriguing because it suggests the agent has learned a strategy that requires no persistent internal information, but rather achieves clock-like functionality using only its trained feedforward weights and external input. In order to characterize the strategies the agents developed and how they differed from one another, we evaluated the trajectories of the agents in action space.

The use of highly controlled stimuli and actions in our task, inspired by neuroscience literature, allows us to perform this analysis in a straightforward way. Because our stimuli were presented on a 2D screen and the only allowed actions were shifts in gaze direction, we could simply analyze the agent’s gaze aligned to moments of interest in trials of each sample interval duration. In Fig. 6 we show this analysis. In the recurrent agent, action trajectories appear similar across the range of sample intervals: the agent maintains fixation in a small region near the initial fixation point throughout the Ready-Set interval (which it must measure), then it linearly shifts gaze to align with the target at the desired time.

On the other hand, the feedforward agent shows a more interesting pattern: after the Ready cue, it begins to traverse a stereotyped trajectory. When the Set cue arrives, it deviates off the trajectory and proceeds along another stereotyped trajectory, which it follows until it reaches the target. By expanding the extent of these trajectories in a consistent way, the agent measures elapsed time. This strategy can be described as one that uses the external environment as a clock, which is rational in the absence of any persistent internal states to use as a clock.

One framework that may explain this feedforward agent’s solution has been studied in animal behavior research and is called *stigmergy* (26): coordination with the external environment to indirectly transfer information across individuals. In this case, one may describe the stereotyped action pattern used for interval timing as “autostigmergy”: the agent’s own interactions with the environment serve as a source of memory external to the agent, but which can nevertheless be used to guide its actions. This “autostigmergic” solution is a particularly interesting proposition in light of the existing literature on mechanisms of timing in animals: many studies have suggested that animals may measure and encode time through a process inherently linked with their behavior (16). In other words, rather than explicitly implementing a clock using neural activity, they indirectly measure time through transitions in behavioral space. Fascinatingly, in a recent study where rats were trained to time out a particular interval of time, the authors demonstrated that the rats solved the task by developing highly stereotyped movements that spanned the target interval, suggesting a possible link to behavioral theories of timing (15). The stigmergic strategy we observed here with our agents is therefore similar in nature to the strategy rats naturally adopted in that study. This observation suggests that animals and artificial agents may converge on similar solutions for interval timing, and this may be a consequence of shared computational constraints across both systems.

### 3.5 Exploring architectural variants

Given the difference between the behaviors of the recurrent and feedforward agents, we explored and compared the performance of some alternative architectural variants (Fig.7). The goal of this analysis was to briefly explore the sensitivity of the agent’s performance to its specific setup, and future work should extend these analyses to deeper characterization of a broader range of architectures.

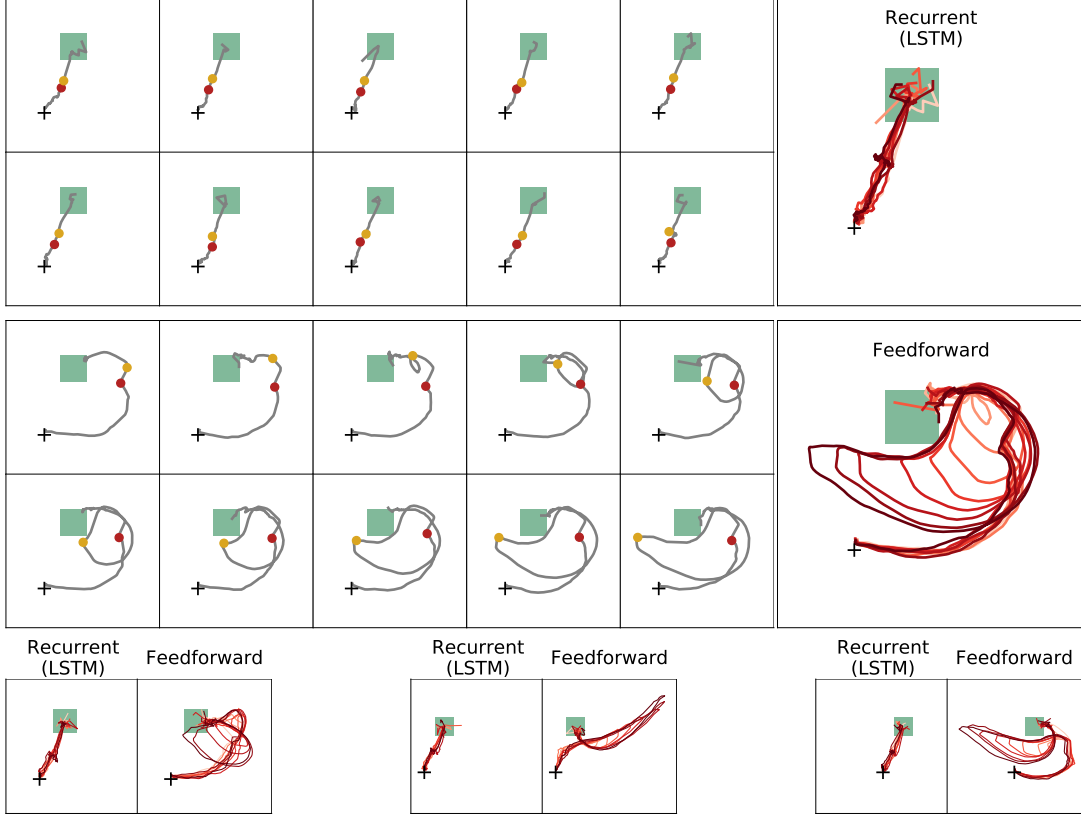


Figure 6: Agent gaze trajectories. (a) The gaze position of the agent was recorded at each time point throughout the trial for rewarded trials of varying sample interval durations. Each panel shows a schematic of the environment, with the black cross representing the central trial initiation gaze target, and the green square representing the Go target. Each subpanel shows the trajectory of gaze position over time (gray line) averaged across trials of one specific sample interval duration. For reference, the red dot corresponds to the moment when the “Ready” cue appeared, and the yellow dot to when the “Set” cue appeared. The upper left panel shows the mean over trials with the shortest sample interval, increasing rightward, with the bottom right showing the longest. The large right-side panel shows the trajectories overlaid, colored according to sample interval duration (the darkest red corresponds to the longest sample interval, i.e. the trajectory in the bottom-right small panel). (b) The same as shown in a, but for the feedforward agent. (c) Three more pairs of examples from different seeds comparing the recurrent and feedforward agents.

We first varied the number of hidden units (set at 128 throughout the study) to determine whether fewer parameters in the LSTM agent might degrade performance, or whether more parameters in the feedforward agent may augment performance. These alterations had minimal effect on the overall performance of the agents. We next trained agents in which the controller was not an LSTM or feedforward network but rather a vanilla RNN, GRU (3) or RMC (relational memory core) (24) instead. Agents with these controllers all learned the task, though the vanilla RNN and RMC exhibited some biases in performance. We proceeded to ask about the performance of a frozen LSTM: that is, its parameters were non-trainable; they were initially randomized and not changed thereafter throughout training. Interestingly, this agent learned to the same degree as the basic LSTM agent. This finding aligns with other reports that learning can occur in networks with random weights (18). Given this apparent robustness, we then asked whether it would learn when the reinforcement learning algorithm was limited to policy and baseline updates on smaller segments of agent-environment interactions (i.e. fewer steps). In particular, we modified the agent such that the 100 steps used in backpropagation through time were divided into 10 chunks (of 10 steps each) for the sake of computing the policy gradient and baseline losses for the reinforcement learning algorithm (See A.1.1 for details). In this truncated “10-step RL,” the reinforcement learning algorithm trained on episodes far shorter

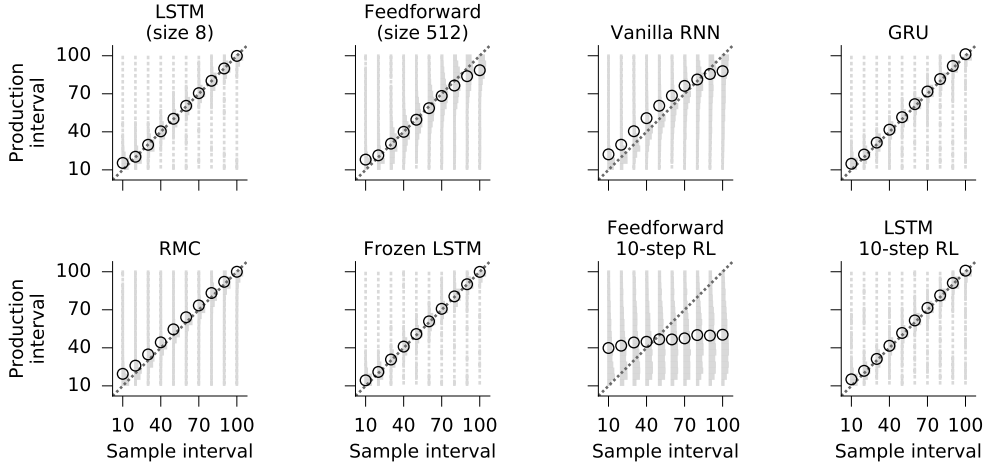


Figure 7: Performance with architectural variants. Display conventions are as in Fig. 3b. LSTM/Feedforward size indicates the number of hidden units in the controller (as compared to 128 in all prior figures). Vanilla RNN, GRU, and RMC were substituted in as replacements for the LSTM/feedforward controllers. Frozen LSTM is an LSTM controller where parameters are not trainable. 10-step RL refers to an agent that uses chunks of 10 agent-environment steps to compute policy- and baseline-gradient updates (as compared to 100 steps in all prior figures).

than the temporal intervals to be learned. We found that while the feedforward agent (which lacks backpropagation through time) was severely impaired by this alteration, the recurrent agent was not.

## 4 Discussion

Here we adapted an interval timing task from the neuroscience literature and used it to study deep reinforcement learning agents. We found that both recurrent and feedforward agents could solve the task in an end-to-end manner. We furthermore characterized differences in the behaviors of the agents at the levels of timing precision and generalization, hidden unit activations, and trajectories through action space. Recurrent agents implemented timers that could be characterized as counters in the LSTM hidden units, whereas feedforward agents developed stigmergy-like strategies that bear resemblance to psychophysical results from timing experiments in animals.

The importance of understanding interval timing in deep reinforcement learning agents has been previously recognized (23), and other work has been performed using neural networks to study time perception. For example, (25) showed that neural networks trained on time estimation from visual scenes can be used to model specific biases humans exhibit in time perception. In addition to the temporal reproduction task we studied here, there exist other timing tasks that are commonly used in the animal literature, such as temporal production (15) and temporal discrimination (21) tasks. Therefore, we have also generated tasks like this in PsychLab for future study, and we are open-sourcing all these tasks as part of this contribution.

Future work should explore the ways in which different environmental and agent architectural constraints alter the solutions of the agent. Furthermore, it will be useful to determine how findings from interval timing tasks like these, performed in controlled psychology-like environments, generalize to more complex domains where interval timing is necessary but is not the primary goal. Finally, characterizing agents’ solution space for fundamental abilities like timing will be useful in designing future challenges and solutions to more complex tasks for AI. Perhaps the stigmergic behavior we uncovered in this study indicates the broader importance of deeply characterizing – and possibly controlling – agent behaviors in conjunction with their architectures when designing and studying intelligent abilities.



## References

- [1] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. *CoRR*, abs/1612.03801, 2016.
- [2] C. V. Buhusi and W. H. Meck. What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.*, 6(10):755–765, Oct 2005.
- [3] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.
- [4] R. M. Church and M. Z. Deluty. Bisection of temporal intervals. *J Exp Psychol Anim Behav Process*, 3(3):216–228, Jul 1977.
- [5] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. *CoRR*, abs/1802.01561, 2018.
- [6] John Gibbon. Scalar expectancy theory and weber’s law in animal timing. *Psychological Review*, 84:279–325, 03 1977.
- [7] E. Hazeltine, L. L. Helmuth, and R. B. Ivry. Neural mechanisms of timing. *Trends Cogn. Sci. (Regul. Ed.)*, 1(5):163–169, Aug 1997.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [10] Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *arXiv preprint arXiv:1810.06721*, 2018.
- [11] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint arXiv:1807.01281*, 2018.
- [12] M. Jazayeri and M. N. Shadlen. Temporal context calibrates interval timing. *Nat. Neurosci.*, 13(8):1020–1026, Aug 2010.
- [13] Mehrdad Jazayeri and Michael N Shadlen. A neural mechanism for sensing and reproducing a time interval. *Current Biology*, 25(20):2599–2609, 2015.
- [14] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].
- [15] R. Kawai, T. Markman, R. Poddar, R. Ko, A. L. Fantana, A. K. Dhawale, A. R. Kampff, and B. P. Olveczky. Motor cortex is required for learning but not for executing a motor skill. *Neuron*, 86(3):800–812, May 2015.
- [16] P. R. Killeen and J. G. Fetterman. A behavioral theory of timing. *Psychol Rev*, 95(2):274–295, Apr 1988.
- [17] Joel Z Leibo, Cyprien de Masson d’Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018.

- [18] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nat Commun*, 7:13276, 11 2016.
- [19] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.
- [20] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*, 2016.
- [21] S. Pai, J. C. Erlich, C. Kopec, and C. D. Brody. Minimal impairment in a rat model of duration discrimination following excitotoxic lesions of primary auditory and prefrontal cortices. *Front Syst Neurosci*, 5:74, 2011.
- [22] J. J. Paton and D. V. Buonomano. The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, 98(4):687–705, May 2018.
- [23] E. A. Petter, S. J. Gershman, and W. H. Meck. Integrating Models of Interval Timing and Reinforcement Learning. *Trends Cogn. Sci. (Regul. Ed.)*, 22(10):911–922, Oct 2018.
- [24] Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks, 2018.
- [25] Marta Suárez-Pinilla, Kyriacos Nikiforou, Zafeirios Fountas, Anil Seth, and Warrick Roseboom. Perceptual content, not physiological signals, determines perceived duration when viewing dynamic, natural scenes, Dec 2018.
- [26] G. Theraulaz and E. Bonabeau. A brief history of stigmergy. *Artif. Life*, 5(2):97–116, 1999.
- [27] Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M. Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Yuhuai Wu, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. <https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>, 2019.
- [28] Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z Leibo, Adam Santoro, Mevlana Gemici, Malcolm Reynolds, Tim Harley, Josh Abramson, Shakir Mohamed, Danilo Rezende, David Saxton, Adam Cain, Chloe Hillier, David Silver, Koray Kavukcuoglu, Matthew M Botvinick, Demis Hassabis, and Timothy Lillirap. Unsupervised predictive memory in a goal-directed agent. *arXiv preprint arXiv:1803.10760*, 2018.
- [29] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X. J. Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.*, 22(2):297–306, 02 2019.

## A Appendix

### A.1 Importance Weighted Actor-Learner Architecture

Importance Weighted Actor-Learner Architecture (IMPALA) (5) takes advantage of an off-policy actor-critic approach. Decoupled actors in conjunction with an environment communicate experience to a learner worker in a many-to-one (actor-to-learner) relationship. Each actor generates a batched trajectory, or episode, of experience and sends the state-action-reward traces  $(s_0, a_0, r_0, \dots, s_k, a_k, r_k)$  to its assigned learner. Learners gather trajectories from all actors, compute the policy and model gradients and update the model parameters continuously. Actors receive parameter updates from the learner then continue to generate trajectories upon completing assembly of a trajectory.

Actor and learner policies become increasingly unsynchronized between parameter updates. The actor's *behaviour policy*,  $\mu$ , is said to have *policy lag* with respect to the *target policy* of the learner,  $\pi$ . Importance weighting with *V-trace* targets are computed at each step to account for *policy lag*:

$$v_s \stackrel{d}{=} V(x_s) + \sum_{t=s}^{s+n-1} \gamma^{t-s} (\pi_{i=s}^{t-1} c_i) \rho_t (r_t + \gamma V(x_{t+1}) - V(x_t)) \quad (1)$$

where  $\gamma \in [0, 1)$  is a discount factor,  $x_t$  and  $r_t$  are the state reward at time-step  $t$ ,  $\rho_t = \min(\bar{\rho}, \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)})$  and  $c_i = \min(\bar{c}, \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)})$  are truncated importance sampling weights. V-trace targets are then used to compute gradients for the policy approximation in the learner. This has the desirable effect of allowing observations and parameters to flow in a single direction, permitting higher data efficiency and resource allocation in comparison to alternate paradigms such as asynchronous advantageous actor critic (A3C) (19).

#### A.1.1 Chunked RL Horizons

The value trace above can be modified simply to account for k-step chunking described in 3.5. Defining  $K$  even sized chunks where  $n$  is the number of steps in a single chunk the value estimate  $v_s$  is reformulated in Eq. 2.

$$v_s \stackrel{d}{=} V(x_s) + \frac{1}{K} \sum_{k=1}^K \sum_{t=s+n(k-1)}^{s+nk-1} \gamma^{t-s} (\pi_{i=s}^{t-1} c_i) \rho_t (r_t + \gamma V(x_{t+1}) - V(x_t)) \quad (2)$$