Cite as: Jiang, L. P., & Rao, R. P. N. (2023). Dynamic Predictive Coding Explains Both Prediction and Postdiction in Visual Motion Perception. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 45, No. 45).

Dynamic Predictive Coding Explains Both Prediction and Postdiction in Visual Motion Perception

Linxing Preston Jiang (prestonj@cs.washington.edu)
Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195 USA

Rajesh P. N. Rao (rao@cs.washington.edu)
Paul G. Allen School of Computer Science & Engineering and Center for Neurotechnology
University of Washington, Seattle, WA 98195 USA

Abstract

Due to transmission delays, the perceptual information our brain can access quickly becomes outdated as events unfold in real-time. We suggest our perceptual system learns internal representations that encode sequences (or timelines) rather than single points to compensate for transmission delays. Specifically, we investigate the dynamic predictive coding (DPC) model in which high-level states predict the *transition dynamics* of lower-level states and represent lower-level state sequences. We show that a two-level DPC network trained to predict videos captures several aspects of the well-known flash-lag illusion and exhibits both predictive and postdictive effects resembling those observed in human visual motion processing. Our results support the view that visual perception relies on temporally abstracted representations that encode sequences (or timelines) rather than single time points.

Keywords: visual perception; predictive processing; flash-lag illusion; postdiction; apparent motion

Introduction

Sensory systems enable us to perceive and interact with a highly dynamic world in real time. Yet, biological constraints like neural processing delays deny us access to the physical present. How does the nervous system compensate for these delays and create a multisensory percept that synchronizes with the world? Our ability to predict future stimuli and event outcomes seems crucial in solving this problem. Indeed, predictive representations of upcoming stimuli have been found in various open and closed-loop paradigms where animals developed experience-dependent visual and auditory expectations (Xu, Jiang, Poo, & Dan, 2012; Keller, Bonhoeffer, & Hübener, 2012; Gavornik & Bear, 2014; Fiser et al., 2016; Schneider, Sundararajan, & Mooney, 2018). Trajectory extrapolation (prediction) has also been suspected to underlie many visual motion processing phenomena observed in psychophysical studies (Nijhawan, 1994, 2008; Hogendoorn, 2020; Lotter, Kreiman, & Cox, 2020).

However, predictive mechanisms alone fail to explain many reports that perception of earlier sensory information (in a sequence) can sometimes be altered by stimuli that arrive later (Shimojo, 2014). This phenomenon, referred to as "postdiction", challenges the intuitive view that perception strictly follows the order of sensory events. For instance, Eagleman & Sejnowski showed that in the classic flash-lag illusion experiment (Nijhawan, 1994), the direction of the moving object after the flash could change the direction of the perceived displacement of the flash (Eagleman & Sejnowski, 2000). Others have shown that when future events deviate from expectation, predictive and postdictive effects dominate apparent motion perception at different latencies (Hogendoorn, Carlson, & Verstraten, 2008; Blom, Feuerriegel, Johnson, Bode, & Hogendoorn, 2020; Blom, Bode, & Hogendoorn, 2021). The neural mechanism through which new sensory information is incorporated to edit earlier percepts remains unclear.

Here, we adopt the hypothesis that our perceptual system encodes entire sequences rather than single points at any given time (Hogendoorn, 2022). Such a sequence representation allows the system to predict the expected perceptual trajectory, compensating for transmission delays. When future events deviate from this expectation, the system retroactively updates its sequence representation and catches up with new observations. We hypothesize that dynamical illusions such as the flash-lag effect and apparent motion could be explained by these editable "timeline" representations the perceptual system forms.

To test this hypothesis, we studied a neural model called dynamic predictive coding (DPC) which learns hierarchical representations of sequences (Jiang, Gklezakos, & Rao, 2021). In the DPC formulation, lower-level states predict both the current sensory input and the next state, while a higher-level state predicts the *transition dynamics* between lower-level states. This enables higher-level states to predict *entire sequences* of lower-level states following the same dynamics. We demonstrate that a two-level DPC network trained to predict image sequences naturally exhibits the flash-lag effect under different testing conditions (Eagleman & Sejnowski, 2000). Moreover, the sequence prediction and error correction process of DPC explain the observed interplay between prediction and postdiction in apparent motion perception (Hogendoorn et al., 2008). Taken together, these results support the view that visual perception relies on temporally abstracted representations that encode sequences (or timelines) rather than single points, naturally leading to predictive and postdictive effects in visual perception.

Dynamic Predictive Coding

The DPC model assumes that spatiotemporal inputs are generated by a hierarchical generative model (Figure 1a). The lower level of the model follows the traditional predictive coding model in generating images using a set of spatial filters $\mathbf{U}$ and a latent state vector $\mathbf{r}_t$, which is sparse (Olshausen & Field, 1996), for each time step t: $\mathbf{I}_t = \mathbf{U}\mathbf{r}_t + \mathbf{n}$ where $\mathbf{n}$

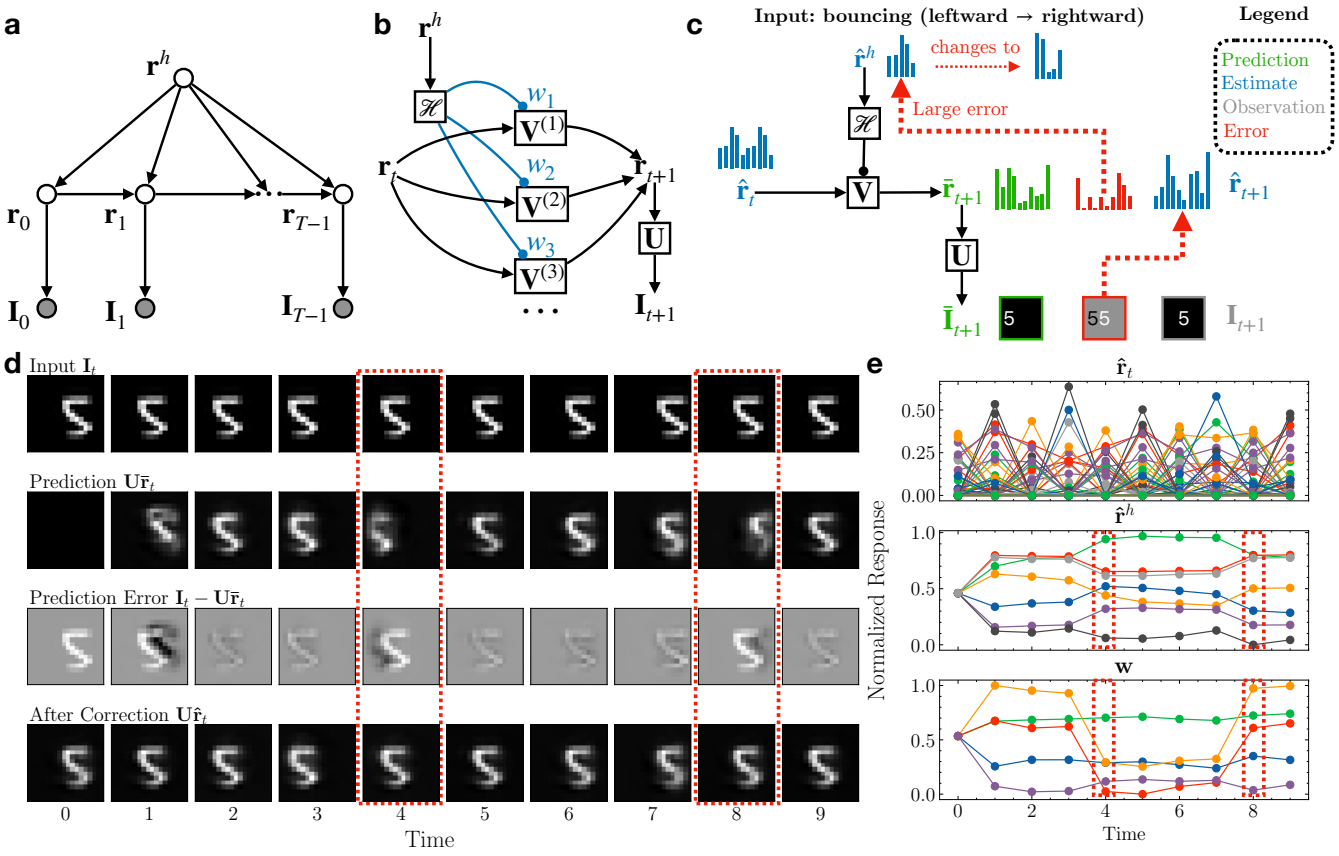

Figure 1: **Dynamic predictive coding.** **(a)** Generative model for dynamic predictive coding. **(b)** Parameterization of the model. The higher-level state modulates the lower-level transition matrices through a top-down network $\mathcal{H}_\theta$. **(c)** Depiction of an inference step when the lower-level dynamics changes. The resulting large prediction errors drive updates to the higher-level state to account for the new lower-level dynamics. **(d)** Inference in a trained network for an example input sequence from the Moving MNIST dataset. The red dashed boxes mark the time steps when the dynamics of the input changed. **(e)** The network's responses to the input Moving MNIST sequence in (d). Note the changes in the higher-level responses after the input dynamics changed (red dashed boxes); this gradient-based change helps to minimize prediction errors.

is zero mean Gaussian white noise. The temporal dynamics of the state $\mathbf{r}_t$ is modeled using K learnable transition matrices $\mathbb{V} = \{\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(K)}\}$ which can be linearly combined using a set of "modulation" weights given by a K-dimensional mixture vector $\mathbf{w}$. This vector of weights is generated by the higher-level state vector $\mathbf{r}^h$ using a function $\mathcal{H}_\theta$ (Figure 1b), implemented as a neural network:

$$\mathbf{w} = \mathcal{H}_\theta(\mathbf{r}^h) \tag{1}$$

$$\mathbf{V} = \sum_{k=1}^{K} w_k \mathbf{V}^{(k)}. \tag{2}$$

Here, w_k is the kth component of the vector $\mathbf{w}$. The lower-level state vector at time $t+1$ is generated as $\mathbf{r}_{t+1} = f(\mathbf{V}\mathbf{r}_t) + \mathbf{m}$ where $\mathbf{m}$ is zero mean Gaussian white noise and f is a nonlinearity function (ReLU in our case).

Inference and Learning

When an input sequence is presented, the model employs a Bayesian filtering approach to perform online inference on the latent vectors by minimizing a loss function that includes prediction errors and penalties from prior distributions over the latent variables:

$$\mathcal{L}_t(\mathbf{r}_t, \mathbf{r}^h, \mathbf{U}, \mathbb{V}, \theta) := \frac{1}{2\sigma^2}\|\mathbf{I}_t - \mathbf{U}\mathbf{r}_t\|_2^2 + \frac{1}{2\sigma_r^2}\|\mathbf{r}_t - f(\mathbf{V}\hat{\mathbf{r}}_{t-1})\|_2^2 + \lambda\|\mathbf{r}_t\|_1 + \lambda_h\|\mathbf{r}^h\|_2^2, \tag{3}$$

where σ^2 is the image noise variance, σ_r^2 is the latent noise variance, λ is the sparsity penalty for $\mathbf{r}_t$, and λ_h is the prior penalty for $\mathbf{r}^h$. $\hat{\mathbf{r}}_{t-1}$ and $\mathbf{r}^h$ are the optimal lower- and higher-level estimates from the previous step, respectively. The optimal estimates for the current step t are then

$$\hat{\mathbf{r}}_t = \arg\min_{\mathbf{r}_t} \mathcal{L}_t(\mathbf{r}_t, \mathbf{r}^h, \mathbf{U}, \mathbb{V}, \theta) \tag{4}$$

$$\hat{\mathbf{r}}^h = \arg\min_{\mathbf{r}^h} \mathcal{L}_t(\mathbf{r}_t, \mathbf{r}^h, \mathbf{U}, \mathbb{V}, \theta). \tag{5}$$

After inferring $\mathbf{r}_t$ for the whole sequence, the parameters are learned by gradient descent on the sum of loss functions from

all steps:

$$\mathcal{L}(\mathbf{U},\mathbb{V},\theta) \coloneqq \sum_{t=1}^{T} \mathcal{L}_t(\hat{\mathbf{r}}_t, \hat{\mathbf{r}}^h, \mathbf{U},\mathbb{V},\theta), \tag{6}$$

where T is the sequence length.

Figure 1c illustrates the inference process for both levels of the network. The network generates top-down and lateral predictions (green) using the current two-level state estimates (blue). If the input sequence is predicted well by the top-down-modulated transition matrix $\mathbf{V}$, the higher-level response $\mathbf{r}^h$ remains stable due to small prediction errors. When a non-smooth transition occurs in the input sequence, the resulting large prediction errors are sent to the higher level via feedforward connections (red arrows, Figure 1c), driving changes in $\mathbf{r}^h$ to predict new dynamics for the lower level.

Dataset and Hyperparameters

We trained a two-level DPC network on the Moving MNIST dataset (Srivastava, Mansimov, & Salakhudinov, 2015). We used 10,000 image sequences (image size: 18×18 pixels, sequence length: $T = 10$ frames), each sequence containing a fixed digit moving in a particular direction. The motion of the digits was restricted to upward, downward, leftward, or rightward directions. When a digit hit the boundary, its motion direction was inverted (leftward to rightward, upward to downward, and vice versa). 9,000 sequences were used to train the model and the remaining 1,000 were reserved for testing.

The DPC network consisted of 648 first-level neurons, 20 second-level neurons, and $K = 5$ first-level transition matrices. The top-down network $\mathcal{H}_\theta$ was a one-hidden-layer multilayer perceptron with 10 hidden units, a LayerNorm layer (Ba, Kiros, & Hinton, 2016) and an ELU activation function (Clevert, Unterthiner, & Hochreiter, 2016). The training process lasted 100 epochs and we used the model weights at the last epoch for the following simulations.

Hierarchical Sequence Representations

To use the trained DPC network for testing our hypothesis, we need to first confirm that higher-level states of the network have developed hierarchical representations that encode entire sequences. Figure 1d illustrates the trained network's inference process on an example image sequence in the test set. As seen in Figure 1e, the lower-level responses displayed fast changes while the higher-level responses spanned a longer timescale and showed greater stability. Note that at time $t = 4$ and $t = 8$, the input dynamics changed as the digit "bounced" against the boundaries and started to move in the opposite motion (Figure 1d red dashed box), inducing large prediction errors at those times (Figure 1d third row). These errors caused notable changes in the higher-level responses $\mathbf{r}^h$ (Figure 1e red dashed boxes). For the rest of the steps, $\mathbf{r}^h$ remained stable and generated accurate predictions of the stable dynamics, coding for the entire leftward or rightward sequences. These results show that the second-level DPC neurons learned more temporally abstract representations that encode entire sequences of lower-level activities following the same transition dynamics.

In the following sections, we show that the ability of the DPC model to encode entire sequences at the higher level (*c.f.* the "timeline" model of perception (Hogendoorn, 2022)) leads to new normative and computational interpretations of visual motion phenomena such as the flash-lag illusion (Nijhawan, 1994; Eagleman & Sejnowski, 2000; Nijhawan, 2008), explaining both predictive and postdictive effects (Hogendoorn et al., 2008; Hogendoorn, 2022). The flash-lag illusion refers to the phenomenon that a flashed, intermittent object is perceived to be "lagged" behind the percept of a continuously moving object even though the physical locations of the two objects are aligned or the same (Nijhawan, 1994, 2008). Though this illusion is commonly attributed to the predictive nature of the perceptual system (Nijhawan, 1994), Eagleman and Sejnowski (2000) proposed a postdictive mechanism based on psychophysical results that the motion of the moving object *after* the flash can change the percept of events at the time of the flash.

We propose that prediction error minimization with a hierarchical temporal representation, as in the DPC model, provides a natural explanation for these predictive and postdictive effects. In a DPC network, the higher-level state $\mathbf{r}^h$ predicts entire sequences of lower-level states following the same dynamics (Figure 1a,c)). When the dynamics of observations change (*e.g.*, motion reversal), the higher-level state is updated to minimize prediction errors, resulting in a revised state that represents the motion-reversed sequence spanning both past and future inputs. This process corresponds to postdiction in visual processing (Shimojo, 2014). For the flash-lag experiment, we predict that the higher-level neurons of a trained DPC network will form a static sequence percept when presented with a flashed object and a directional sequence percept for a moving object, causing perceived lags between the two objects as observed in the flash-lag illusion (Nijhawan, 1994).

Prediction versus Postdiction in the Flash-Lag Illusion

We first test these predictions of the DPC model on the experimental conditions used by Eagleman and Sejnowski (2000). In their experiment, the stimuli consisted of a flashed disk and a ring moving in a circle. Before the flash, the ring could have an initial trajectory (Figure 2a, top) or no initial trajectory (Figure 2a, bottom). After the flash, the ring could continue moving on its initial trajectory ("continuous"), stop moving ("stopped"), or move on the reversed trajectory ("reversed"). A flash appeared in a seven-degree range that extended above and below the ring on its trajectory. The participants then indicated whether a flashed white disk occurred above or below the center of the moving ring. Positive displacements denoted lags along the initial trajectory of the ring, while negative displacements denoted the reversed direction.

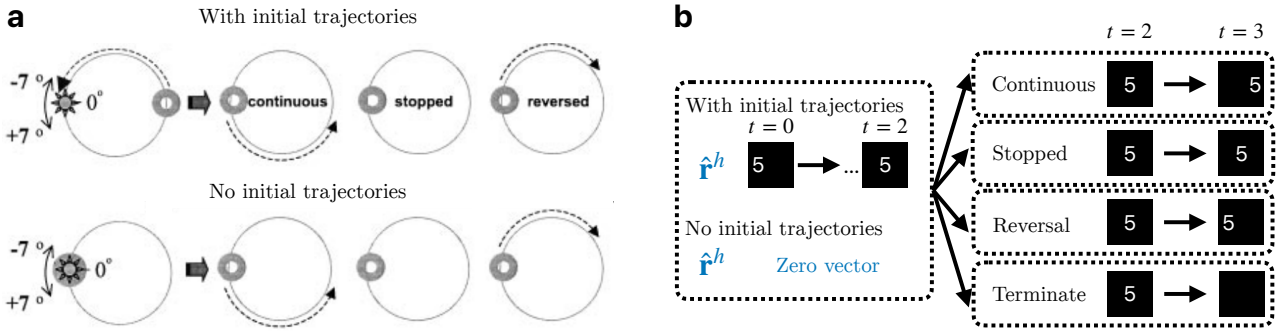

Figure 2: **Flash-lag testing conditions. (a)** The testing conditions used by Eagleman and Sejnowski (2000). The moving ring could have an initial trajectory (top) or no trajectory (bottom). At the time of the flash (bright disk), the ring could move along the initial trajectory, stop, or reverse its trajectory. Adapted from Eagleman and Sejnowski (2000). **(b)** Two test conditions (left) regarding initial trajectories of the moving object (a digit) in the flash-lag experiment with the model, and four test conditions (right) for the moving object. The flashed object was shown at time t and turned off at time $t+1$ (same as the "Terminate" condition).

Simulation

To simulate these testing conditions, we used the Moving MNIST test set and extracted 134 test sequences with consistent leftward or rightward motion. For each of these 134 sequences, we simulated the two test conditions, namely, with or without initial trajectory: the higher-level state $\hat{\mathbf{r}}^h$ was either inferred from the first three steps ($t = 0, 1, 2$) of the input sequence, or initialized to the zero vector (Figure 3(a) left). For each of these two test conditions, we simulated the three test cases regarding the motion of the moving object at the time of the flash (Figure 3(a) right) and an additional "Terminate" case used by Nijhawan (2008). Note that flashed stimuli correspond to the "no initial trajectory, terminate" condition since the model has no belief about the dynamics ($\mathbf{r}^h$ is a zero vector) and the stimuli only appear for one frame.

We computed the location of a digit as the center of mass of pixel values in the 2D image; the perceived location at time t was defined similarly based on the predicted image $\bar{\mathbf{I}}_t$:

$$\bar{\mathbf{w}} = \mathcal{H}_\theta(\hat{\mathbf{r}}^h) \tag{7}$$

$$\bar{\mathbf{V}} = \sum_{k=1}^{K} \bar{w}_k \mathbf{V}^{(k)} \tag{8}$$

$$\bar{\mathbf{I}}_t = \mathbf{U}\left(\mathrm{ReLU}\left(\bar{\mathbf{V}}\hat{\mathbf{r}}_{t-1}\right)\right). \tag{9}$$

Here, $\hat{\mathbf{r}}^h$ is the optimal higher-level estimate at $t-1$ (Equation 5) , and $\hat{\mathbf{r}}_{t-1}$ is the optimal lower-level estimate at $t-1$ (Equation 4). We computed the location of the percept as the center of mass of the percept image $\bar{\mathbf{I}}$. The displacement in percept between the moving object and the flashed object was calculated as

$$\begin{cases} \mathrm{C}\left(\bar{\mathbf{I}}_t^{\text{moving}}\right) - \mathrm{C}\left(\bar{\mathbf{I}}_t^{\text{flash}}\right) & \text{if rightward motion} \\ \mathrm{C}\left(\bar{\mathbf{I}}_t^{\text{flash}}\right) - \mathrm{C}\left(\bar{\mathbf{I}}_t^{\text{moving}}\right) & \text{if leftward motion} \end{cases}, \tag{10}$$

where $\mathrm{C}(\mathbf{I})$ returns the horizontal location of the center of mass of $\mathbf{I}$. Therefore, a positive displacement is along the original trajectory of the moving object, while a negative displacement is along the reversed trajectory.

Results

First, we investigate the model's perception of the flashed object ("no initial trajectory, terminate" case[1]). As Figure 3(c) shows, the perceived location of a flashed object at $t = 3$ ($\bar{\mathbf{I}}_3$) strongly overlapped with the physical flashed location at $t = 2$ ($\mathbf{I}_2$), showing that the prediction errors (induced by the disappearance of the digit) drove the higher-level state estimates to predict no change in object location for the flashed object. Figure 3(d) shows the perceived displacement between the moving object (with initial trajectories) and the flashed object, computed as the difference in perceived locations at $t = 3$ between the moving object ($\bar{\mathbf{I}}_3^{\text{moving}}$) and the flashed object ($\bar{\mathbf{I}}_3^{\text{flash}}$). The perceived displacements in the model (Figure 3(d)) were similar to the psychophysical reports by Eagleman and Sejnowski (2000) in all three test conditions (Figure 3(a)). The perceived displacements in the terminate case were similar to the stopped case (with much less variance), similar to the report by Nijhawan (2008). Figure 3(e) confirms that the initial trajectories of the moving object had no effects on the model's flash-lag illusion, consistent with the reported results (Figure 3(b)) (Eagleman & Sejnowski, 2000).

These results validate the explanation provided by the DPC model on the flash-lag effect: For a hierarchical generative model with representations of sequences, inference on a flashed object or a stopped/terminated moving object leads to a belief of a static object sequence (Figure 3(c)), while continuous or reversed motion leads to a belief of a moving

[1]Note that for any "no initial trajectory" condition, $t = 2$ is the first step and $t = 3$ is the second step.

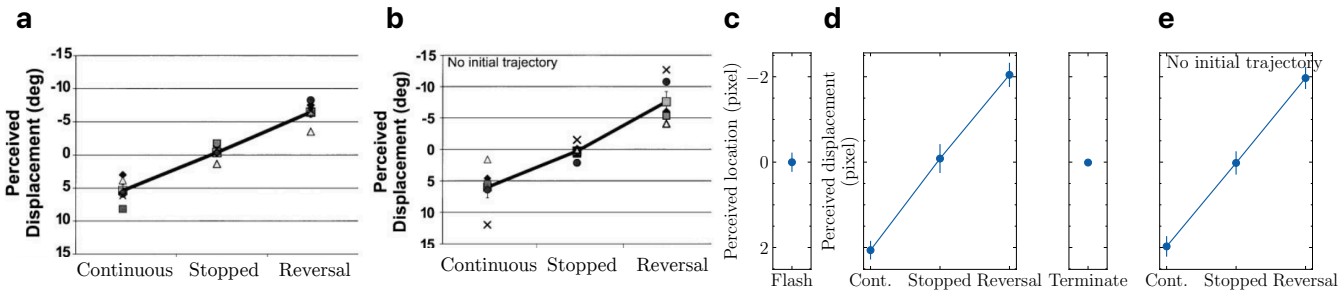

Figure 3: **Flash-lag results. (a) & (b)** The psychophysical estimates for human subjects reported by Eagleman and Sejnowski (2000) when the moving object had initial trajectories (a) or no initial trajectory (b). **(c)** Perceived location of the flashed object at time $t+1$. The error bar indicates ± 1 standard deviation (measured across presentations of different digits). **(d)** The perceived displacement between the moving object (with initial trajectories) and the flashed object for the four test conditions. **(e)** Same as (d) but with no initial trajectory for the moving objects.

object sequence, resulting in the perceived lags (or no-lags) along the corresponding directions.

Apparent Motion Perception

The explanation of the flash-lag effect relies on the second level of the DPC network to minimize the prediction error on the first-level transition dynamics (*e.g.* errors induced by flashed stimuli or unpredictable motion reversal). That is, new sensory information induces prediction errors that "postdictively" update the higher-level sequence representations of the DPC network. One aspect of motion perception the previous results do not illustrate is the interplay between postdiction and prediction. Hogendoorn et al. investigated this effect in an apparent motion perception experiment. Participants were instructed to either report the detection of a visual cue (short latency) or differentiate between two visual cues (long latency) during apparent motion. These visual cues could either be along the apparent motion trajectory or the reversed trajectory. The authors found that upon reversing the apparent motion trajectory, predictive effects dominated perception at short latency (detection task, Figure 4b), with the most interference (measured in terms of the participants' reaction times) along the original trajectory. At longer latency (differentiation task, Figure 4c), most interference was along the reversed trajectories, indicating dominating postdictive effects.

We hypothesize that the prediction error minimization process of DPC could explain this interplay between prediction and postdiction, as illustrated by Figure 4(a). Figure 4a can be seen as a depiction of the gradient-descent-based optimization process of Equation 5 (and Figure 1c). Early percepts of the model are dominated by the spatiotemporal prediction using the optimal estimates from the previous step (Figure 4a left). When a motion reversal occurs, feedforward prediction errors gradually correct the second-level states (Figure 4a middle) until convergence (Figure 4a right). Therefore, late percepts correspond to error-corrected spatiotemporal predictions. Note that due to the discrete nature of the DPC model (unit time steps), this process is considered to happen "at" one time step (*e.g.* early versus late percept "at" $t = 3$ illustrated in Figure 4a).

Results

To test this hypothesis, we used the same trained DPC network and probed its percept of the moving object at the time of reversal under the "with initial trajectory, reversal" condition (Figure 2b). At short latency (10% of steps into prediction error correction, Figure 4a early percept), the perceived locations for the moving object in most test sequences were along the original trajectory, as denoted by positive displacements compared to the final step before reversal ($t = 2$) (Figure 4d blue)). At longer latency (90%, Figure 4a late percept), the moving object's perceived locations were flipped and along the reversed trajectory (negative displacements; Figure 4d green). This is consistent with psychophysical findings (Hogendoorn et al., 2008; Hogendoorn, 2022) that when the motion of the object unexpectedly reversed, prediction effects were observed at short latency ($\approx$ 350 ms, Figure 4b right panel, bright color denotes locations of interference due to prediction) while postdiction effects were observed at longer latency ($\approx$ 620 ms, Figure 4c right panel, bright color denotes locations of interference due to postdiction). Figure 4e plots the moving object's perceived location in our model throughout the error correction process (Figure 4a, all green percepts): the perceived location varies smoothly from being along the original direction initially to along the reversed direction at greater latencies. These results make a testable prediction: if probed at an intermediate level of latency (between 350 ms and 620 ms), the maximal interference should overlap with the object's location at the time of reversal (i.e., at the black dots in Figure 4b,c), as suggested by Figure 3e.

Discussion

Previous normative models of postdiction in visual processing often relate the effect to the concept of Bayesian smoothing (or backward message passing) (Eagleman & Sejnowski,

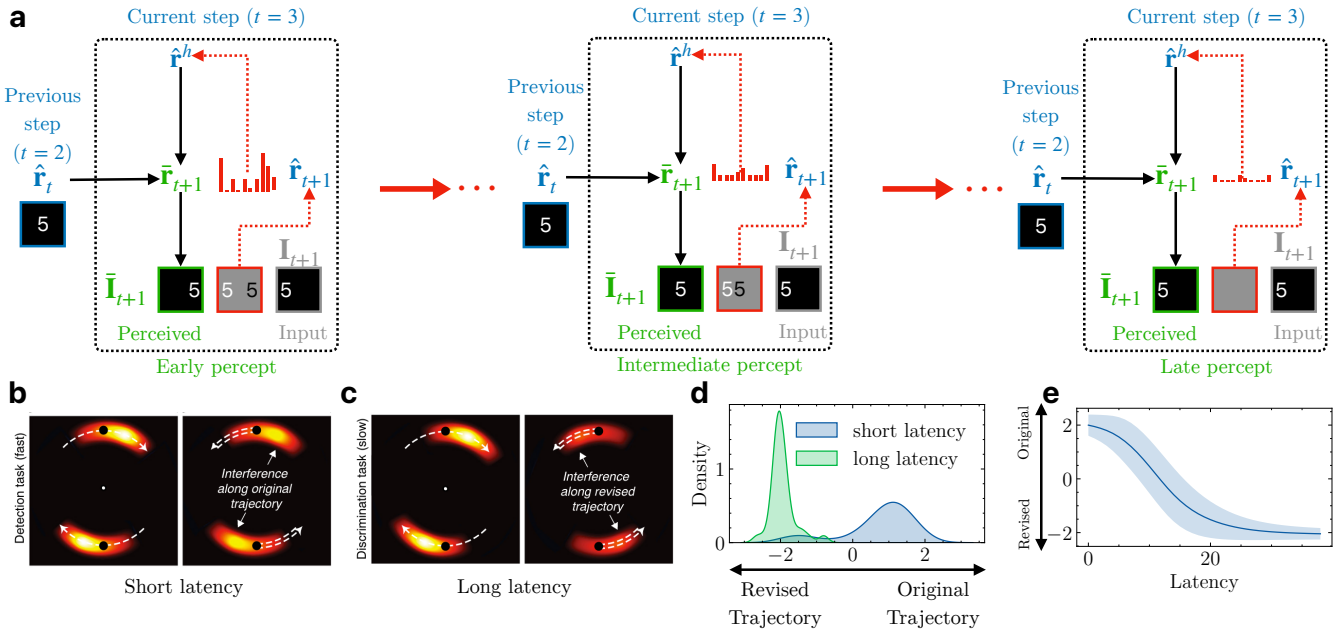

Figure 4: **Predictive versus postdictive effects in apparent motion perception. (a)** Illustration of the prediction-error-driven dynamics of the perception of the moving object when the trajectory reversed at time $t = 3$. Red ellipsis denote the prediction error minimization process. The color legend follows Figure 1c. **(b)** Interference pattern during apparent motion perception with continuous motion (left) and reversed motion (right) at short latency (fast detection task). Brighter color denotes more interference. Dashed arrows represent object motion direction. Adapted from Hogendoorn (2022). **(c)** Same as (b) but at longer latency (slow discrimination task). Adapted from Hogendoorn (2022). **(d)** Perceived location of the moving object at time $t = 3$ probed at short versus long latency during prediction error minimization. Positive values denote distance along the original trajectory. Negative values denote distance along the reversed trajectory. Short and long latency correspond to "Early" and "Late" perception in part (a), respectively. **(e)** Perceived location of the digit at all latencies during the prediction error minimization process in part (a).

2000; Rao, Eagleman, & Sejnowski, 2001). We have shown that a trained two-level DPC network with higher-level sequence representations also exhibits postdictive effects without the need for smoothing. In the event of a temporal irregularity (*e.g.*, an unexpected motion reversal), the higher-level state in the DPC network is updated to reflect a new input sequence, naturally implementing postdiction through online hierarchical Bayesian filtering (Figure 1). Our flash-lag simulation results are consistent with the Bayesian filtering model from Khoei et al. (Khoei, Masson, & Perrinet, 2017) showing that the flash-lag effect can be produced through an internal model that explicitly represents object velocity. The higher-level sequence representation in the DPC model supports an implicit (and more generalized) representation of velocity and reproduces the same internal dynamics of the "speed" estimate at motion reversal (compare Figure 4e with Figure 6 in Khoei et al. (2017)). Such a representation could explain reports that the magnitude of the lag depends on the velocity of the moving objects after the flash (Brenner & Smeets, 2000), as the higher level of the DPC network infers a new dynamics (velocity) estimate through prediction errors induced by the velocity change, which in turn causes the moving object to be perceived at a different distance from the flashed object. It is worth noting that the trained DPC network learned to predict no motion (static sequence) for the flashed object even though it was never trained on static object sequences and did not assume a prior of zero speed (Khoei et al., 2017). This emergent property was also seen in PredNet, which learned to predict relatively little motion for a flashed bar stimulus (Lotter et al., 2020).

In summary, our results demonstrate that the DPC model exhibits predictive and postdictive effects similar to those reported in visual motion processing in humans. DPC unifies the temporal averaging and postdiction models of the flash-lag effect (Hogendoorn, 2020) through temporally abstracted representations of sequences. Directions worthy of further study include (1) more rigorous formulations of the error-correction "time" assumed by DPC and its relation to the window length for temporal averaging (Krekelberg & Lappe, 1999), (2) relaxing the discrete time step assumptions of the generative model, (3) deeper hierarchical versions of the DPC model (Pöppel, 1997; Singhal & Srinivasan, 2021) and (4) the integration of hierarchical actions (Gklezakos & Rao, 2022).

Acknowledgements

LPJ thanks Daogao Liu for inspiring discussions. This material is based upon work supported by National Institutes of Health (NIH) eTeamBCP U01 grant no. 1UF1NS126485-01, National Science Foundation (NSF) EFRI grant no. 2223495, DARPA grant no. HR001120C0021, a Weill Neurohub Investigator grant, and a "Frameworks" grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the funders.

References

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016, July). *Layer Normalization.* ArXiv.

Blom, T., Bode, S., & Hogendoorn, H. (2021, May). The time-course of prediction formation and revision in human visual motion processing. *Cortex*, *138*, 191–202. doi: 10.1016/j.cortex.2021.02.008

Blom, T., Feuerriegel, D., Johnson, P., Bode, S., & Hogendoorn, H. (2020, March). Predictions drive neural representations of visual events ahead of incoming sensory information. *Proceedings of the National Academy of Sciences*, *117*(13), 7510–7515. doi: 10.1073/pnas.1917777117

Brenner, E., & Smeets, J. B. J. (2000, June). Motion extrapolation is not responsible for the flash–lag effect. *Vision Research*, *40*(13), 1645–1648. doi: 10.1016/S0042-6989(00)00067-5

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations.*

Eagleman, D. M., & Sejnowski, T. J. (2000, March). Motion Integration and Postdiction in Visual Awareness. *Science*, *287*(5460), 2036–2038. doi: 10.1126/science.287.5460.2036

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., & Keller, G. B. (2016, December). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, *19*(12), 1658–1664. doi: 10.1038/nn.4385

Gavornik, J. P., & Bear, M. F. (2014, May). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, *17*(5), 732–737. doi: 10.1038/nn.3683

Gklezakos, D. C., & Rao, R. P. N. (2022, January). *Active Predictive Coding Networks: A Neural Solution to the Problem of Learning Reference Frames and Part-Whole Hierarchies.* ArXiv.

Hogendoorn, H. (2020, July). Motion Extrapolation in Visual Processing: Lessons from 25 Years of Flash-Lag Debate. *Journal of Neuroscience*, *40*(30), 5698–5705. doi: 10.1523/JNEUROSCI.0275-20.2020

Hogendoorn, H. (2022, February). Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences*, *26*(2), 128–141. doi: 10.1016/j.tics.2021.11.003

Hogendoorn, H., Carlson, T. A., & Verstraten, F. A. J. (2008, March). Interpolation and extrapolation on the path of apparent motion. *Vision Research*, *48*(7), 872–881. doi: 10.1016/j.visres.2007.12.019

Jiang, L. P., Gklezakos, D. C., & Rao, R. P. N. (2021, February). Dynamic Predictive Coding with Hypernetworks. *bioRxiv*, 2021.02.22.432194. doi: 10.1101/2021.02.22.432194

Keller, G., Bonhoeffer, T., & Hübener, M. (2012, June). Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron*, *74*(5), 809–815. doi: 10.1016/j.neuron.2012.03.040

Khoei, M. A., Masson, G. S., & Perrinet, L. U. (2017, January). The Flash-Lag Effect as a Motion-Based Predictive Shift. *PLOS Computational Biology*, *13*(1), e1005068. doi: 10.1371/journal.pcbi.1005068

Krekelberg, B., & Lappe, M. (1999, August). Temporal recruitment along the trajectory of moving objects and the perception of position. *Vision Research*, *39*(16), 2669–2679. doi: 10.1016/S0042-6989(98)00287-9

Lotter, W., Kreiman, G., & Cox, D. (2020, April). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, *2*(4), 210–219. doi: 10.1038/s42256-020-0170-9

Nijhawan, R. (1994, July). Motion extrapolation in catching. *Nature*, *370*(6487), 256–257. doi: 10.1038/370256b0

Nijhawan, R. (2008, April). Visual prediction: Psychophysics and neurophysiology of compensation for time delays. *Behavioral and Brain Sciences*, *31*(2), 179–198. doi: 10.1017/S0140525X08003804

Olshausen, B. A., & Field, D. J. (1996, June). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. doi: 10.1038/381607a0

Pöppel, E. (1997, May). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, *1*(2), 56–61. doi: 10.1016/S1364-6613(97)01008-5

Rao, R. P. N., Eagleman, D. M., & Sejnowski, T. J. (2001, June). Optimal Smoothing in Visual Motion Perception. *Neural Computation*, *13*(6), 1243–1253. doi: 10.1162/08997660152002843

Schneider, D. M., Sundararajan, J., & Mooney, R. (2018, September). A cortical filter that learns to suppress the acoustic consequences of movement. *Nature*, *561*(7723), 391–395. doi: 10.1038/s41586-018-0520-5

Shimojo, S. (2014). Postdiction: its implications on visual awareness, hindsight, and sense of agency. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00196

Singhal, I., & Srinivasan, N. (2021, December). Time and time again: a multi-scale hierarchical framework for time-consciousness and timing of cognition. *Neuroscience of Consciousness*, *2021*(2), niab020. doi: 10.1093/nc/niab020

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015, June). Unsupervised Learning of Video Representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 843–852).

Xu, S., Jiang, W., Poo, M.-m., & Dan, Y. (2012, March). Activity recall in a visual cortical ensemble. *Nature Neuroscience*, *15*(3), 449–455. doi: 10.1038/nn.3036