Winter 12-9-2023

Predictive Looking and Predictive Looking Errors in Everyday Activities

Sophie Su
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological and Brain Sciences

Predictive Looking and Predictive Looking Errors in Everyday Activities

by
Sophie Su

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

December 2023
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

I would like to extend my deepest gratitude to my advisor Dr. Jeff Zacks for his invaluable guidance, patience, and insightful feedback throughout my research journey. His expertise and support have been instrumental in shaping both this thesis and my academic development.I would also like to thank Dr. Zachariah Reagh and Dr.Richard Abrams for serving on my committee.

Special thanks are also due to my mother, whose unwavering support and encouragement have been a constant source of strength and inspiration. Her help in navigating the challenges of this project has been immeasurable.

Sophie Su

Washington University in St.@ Louis December 2023

Dedicated to my parents.

ABSTRACT OF THE THESIS

Predictive Looking and Predictive Looking Errors in Everyday Activities

by

Sophie Su

Master of Arts in Psychological and Brain Sciences

Department of Psychological and Brain Sciences

Washington University in St. Louis, 2023

Professor Jeffrey Zacks, Chair

People spontaneously segment continuous streams of experiences into distinct episodes. Prediction errors are theorized to drive segmentation. However, existing studies exploring the relationship between prediction error and event segmentation lack a continuous measure of prediction error during naturalistic perception and often fail to distinguish between prediction error and prediction uncertainty. Do moment-by-moment fluctuation of prediction errors, not prediction uncertainty during naturalistic event perception correlate positively with segmentation probabilities? To tackle this question, we harnessed the predictive nature of eye movements and introduced a predictive looking model. In this model, individuals' prior gaze patterns act as predictors for critical features in the current frame. Testing the model using group gaze density maps from participants engaged in passive movie viewing—with actors' hand locations serving as an approximation for the prediction target—we uncovered that past gaze patterns, up to 9 seconds prior, predict the current locations of actors' hands in the movie. Furthermore, a significant and positive correlation emerged between predictive looking errors and prediction errors generated by a computational model. This suggests a congruence in capturing true prediction error signals in the brain. Crucially, aligning with theories proposing an association between increased prediction errors and event segmentation, predictive looking errors positively correlated with event segmentation probabilities.

Chapter 1: Introduction

How can humans effortlessly navigate the continuous, complex, and uncertain world around them? Previous research has highlighted two key strategies: generating predictions of imminent futures and segmenting the ongoing streams of daily experiences into discrete, meaningful events. The question arises: are these two strategies interconnected? Event Segmentation Theory (EST[1]) posits that an increase in prediction error—defined as the discrepancy between prediction and reality—leads to segmentation. A computational model of event comprehension that monitors prediction error and segments when these errors increase has been shown to align with human segmentation, forming human-like event categories [2]. To test this theory of event perception and validate the computational model, our study quantified the extent of predictive looking and prediction errors during naturalistic movie viewing. We further examined the relationships between predictive looking errors, computational model-based prediction errors, and segmentation probabilities.

1.1 The Role of Prediction and Prediction Error in Cognition

Anticipate threats and opportunities before they become imminent is a pivotal strategy for survival in an uncertain and complex world. Prediction has been identified to be crucial for numerous psychological processes. For instance, in scene perception, where people look and attend to are based on their predictions of where goal-relevant objects are likely to be found. [3–8]. Within the domain of action control, it is theorized that predicted outcomes, such as anticipating the green light while crossing the street, are filtered from pedestrians' perception. This filtering mechanism enables individuals to allocate cognitive resources, such as attention and control, to unexpected events that hold greater behavioral relevance, such as unexpected cars running through a red light.

In the realm of learning, updating based on predictions of action values often leads to optimal

strategies.[9, 10]. These empirical findings align with the predictive brain hypothesis, which posits that the brain is not merely a reactive mechanism responding to external stimuli, but rather a proactive system that formulates hypotheses, anticipates the consequences of our actions, and constructs expectations [11–13].

Prediction error plays a pivotal role in updating our internal representation of the environment, with one well-understood form being the temporal difference error in reinforcement learning. This type of prediction error enables us to adjust value we assign to stimuli and outcomes of our actions, ultimately to optimize our actions. For example, as a person starts to learn to play the piano, they may have a prediction of how each combination of notes should sound. Through practicing, they encounter prediction errors, discrepancies between their expected sound and the actual sound produced. Through repeated practice and feedback, the individual gradually refines their mental model, reducing the temporal difference errors over time, leading to improved performance and mastery of the musical piece. At the implementational level, midbrain dopaminergic neurons exhibiting heightened activity in response to predictive cues. The intensity of this activity reflects the magnitude, likelihood, and temporal delay of the anticipated reward [14].

1.2 Event segmentation May Reflect Predictive Processing

Prediction error is also proposed to be a crucial factor in shaping our perception of daily experiences. According to EST; [1] individuals continually generate predictions about what will occur next during ongoing perception. These predictions are thought to stem from internal models known as working event models, known as working event models, which combine sensory input with knowledge of event types (schemata). When the current working event model fails to accurately represent the ongoing situation, resulting in an increase in prediction error, internal working event model will be updated. Subjectively, individuals experience an event boundary between the preceding and current event models. In controlled laboratory settings, with little training participants can identify event boundaries by pressing a button whenever they believe one natural and meaningful unit of activity

has ended and another has begun while observing the activity [15]. Segmentation judgments are reliable across and within participants [16, 17]. Furthermore, when asked to identify events of different grains, segmentation is hierarchically organized such that shorter, fine-grained events are nested within longer, coarse-grained events [15]. Neuroimaging studies showed that during movie viewing, cortical patterns shift systematically and align with participants' marked event boundaries, suggesting that event boundaries are a normal aspect of ongoing perception [18].

1.3 Computational and Behavior Evidence for Prediction Error Driven Segmentation

The hypothesis that an increase in prediction errors leads to the emergence of event boundaries is supported by both behavioral studies and computational simulations. For instance, by pausing the video and directly asking participants for their predictions, Zacks et al., [15] found that people's predictions of what would happen five seconds later are more prone to error around event boundaries compared to that around the middle of events. One limitation of this method is that, by pausing the movie intermittently, researchers inevitably disrupted people's online perception and altered how people approach the viewing task.

Computational modeling provides a valuable avenue for simulating ongoing human perception and testing theories pertaining to the mechanisms underlying event perception updates. One such model, SEM-PE, integrates a recurrent neural network for short-term dynamics with Bayesian inference over event types for seamless event-to-event transitions. Employing learned event schemas in conjunction with observed perceptual information, this architecture constructs a series of event models. This computational model predicts activity dynamics base on the event model, monitors the errors of prediction based on the working event model and updates the current working event model when there is an increase in prediction error. Trained through a single pass of an 18-hour corpus of naturalistic human activity, this model demonstrated its proficiency when tested on an

additional 3.5 hours of activities. SEM-PE successfully segments activities in a manner reminiscent of human perception and forms event categories comparable to those identified by humans [2].

Computational models also offer unique opportunities to differentiate signals that are correlated in natural stimuli by adequately operationalizing continuous internal processes in the brain. In the realm of event cognition, another measure of the dynamics of prediction is prediction uncertainty. Alternative theory of event cognition proposes that it might be the increase in prediction uncertainty, especially toward the end of goals [19, 20] drives event segmentation. Although prediction errors and prediction uncertainties correlate in naturalistic activities, they represent distinct signals for model updating. Prediction error measures the discrepancy between a model's predicted output and the observed ground truth, serving as a metric for how accurately the model's predictions align with the actual data. Prediction uncertainty, on the other hand, reflects a lack of confidence or knowledge about a prediction, indicating the model's level of uncertainty regarding its own predictions.

To investigate the specificity of a prediction error measure, SEM-UNCERTAINTY, a computational model of event cognition identical to SE-PE but with different updating signals, can be employed. This allows an examination of the relative appropriateness of the measure by comparing its fitness to SEM-PE's prediction error against SEM-UNCERTAINTY's prediction uncertainty. A specific measure of prediction error should align more closely with model-generated prediction errors rather than uncertainties.

1.4 The Necessity of Examining the Updating Signals of Proposed Mechanisms

Merely showcasing that a hypothesis-inspired computational model can replicate human-like behaviors is a necessary step but falls short of being sufficient to validate the underlying theory. Consider the domain of speech recognition, where automatic speech recognition (ASR) systems have made significant strides, becoming ubiquitous tools employed by millions globally. While both

human and machine speech recognition systems can be framed as Bayesian perceptual inference processes—with speech as sensory input and the most likely word sequence as the desired output [21, 22]—notable distinctions exist in how machines and humans perceive voice patterns.

Unlike the hierarchical structure of human speech recognition process [23], computational speech recognition systems don't break voice recognition process into discrete parts, such as determining likely speech segments given the sounds heard and identifying likely words given these segments. Instead, computational speech recognition systems integrate acoustic and higher-level information in a single, integrated search process, making the machine agnostic to which speech segments were heard, but only those specific words. Furthermore, human speech recognition encompasses non-acoustic features like lip, mouth, and tongue movements, which are integral to producing speech sounds. However, these features are typically not incorporated into the machine's voice recognition process. Consequently, beyond evaluating the parallels between a computational model's final output and human behaviors, a comprehensive understanding of intermediate outputs—the updating signals of computational models—becomes crucial.

To claim that a prediction error updating computational model of event comprehension is modeling how human segment and comprehend everyday activities, one of the necessary conditions is that model's prediction error correlates well with a measure of human's internals prediction error during passive movie viewing. To examine this condition, a continuous and covert measure of human prediction error during movie viewing is necessary.

1.5 Gaze Reveals Human's Predictive Processes

Eye tracking emerges as a valuable tool for continually and covertly measuring human prediction error. In the realm of statistical learning, researchers assess infants' learning through anticipatory eye movements directed towards locations where stimuli are expected probabilistically[24]. Similarly, in sports, athletes demonstrate predictive behavior by looking ahead to the expected position of a ball rather than simply tracking its current location [3]. Within complex scenes, an individual's

focus of attention is guided by predictions about the most meaningful and task-relevant information present [4].

Eye tracking can also measure the dynamics of human prediction during the viewing of naturalistic activities [3, 4]. Eisenberg and colleagues [25] developed the Predictive Looking at Action Task (PLAT) to investigate the time course of predictability during video watching. Predictive looking was quantified as the duration participants fixated on the object to be touched during successive 500-ms time windows in the three seconds preceding the actual contact. The results revealed that viewers predict and direct their gaze ahead to objects about to be contacted. Moreover, around event boundaries, predictive looking was delayed compared to predictive looking in the middle of an event. This suggests that near an event boundary, predicting future actions becomes more challenging, leading to increased looking times just before contact. One limitation of the PLAT is its limited measurements of prediction error in the time preceding an object contact, leading to sparse sampling of prediction error. Moreover, prediction uncertainty confounds with prediction error, as increase in prediction uncertainty could also results in delay in predictive looking around event boundaries. Consequently, it remains unclear whether individuals engage in continuous predictive looking throughout ongoing activities and whether moment-to-moment prediction error can be accurately derived based on their looking pattern.

1.6 The Current Study

There were five goals for this study: 1) Develop a covert and continuous measure of predictive looking error; 2) Quantify the temporal dynamics of predictive looking during passive movie viewing; 3) Compare this biological measure of prediction error with prediction error generated by computational models; 4) Replicate previously established findings regarding the consistency of observers' judgments of event boundaries; and 5) Test the hypothesis that an increase in prediction error leads to event segmentation.

To address these goals, participants' eye movements were tracked while they passively viewed

video recordings of different solo actors engaging in everyday activities. Building upon prior research indicating that individuals look predictively at an actor's hand [25], predictive looking was operationalized as gaze directed toward locations where the actor's hands would be in the next few seconds. This definition was formalized using a mixed-effect logistic regression model, incorporating both current and past gaze values as predictors of the presence of the actor's hands. To quantify the extent of predictive looking, a model comparison approach was employed, wherein a more complex model included gaze values from further in the past. Prediction error was quantified as the magnitude of the deviance residual of the best-fitted predictive looking model. This biological measure was then compared with two updating signals from two computational models of event comprehension. To replicate previous findings concerning the consistency of individuals' event boundary judgments, an independent sample of participants was recruited to segment the same movies into coarse or fine-grained events. Finally, to explore the relationship between this measure of prediction error and event boundaries, segmentation probabilities were derived based on participants' segmentation data and compared with predictive looking errors.

Chapter 2: Methods

2.1 Eye Tracking

2.1.1 Participants

For this part of the study, we recruited participants from the subject pool at Washington University. In exchange for their time, participants were offered compensation in the form of either course credit or $10 per hour. A target sample size of 100 was decided before data collection to ensure adequate to detect small to middle effect size reported in previous literature [26]. Based on the power analysis, 70 participants were necessary for power of .80. Because some participants might not have usable eye-tracking data, we decided to recruit up to 100 participants. Data from 13 participants were excluded because of experiment program failure (n=7), eye-tracking calibration failure (n=5) or failure to remain on the headrest throughout the study (n=1).

A total of 87 participants successfully completed the study, with their ages ranging from 18 to 32 years (mean age = 19.9 years). The gender distribution of the participants included 28 males and 59 females.

2.1.2 Materials

Four movies of actors performing everyday activities were selected from the META corpus (Bezdek et al., 2022). This stimulus set includes different actors exercising (586 s), making breakfast (586 s), grooming (646 s), and tidying up the room (679 s). The movies were filmed from a fixed, head-height perspective, with no pan or zoom. The frame rate was 60 frames per second, and the frame's dimension was 1280*960. This movie set can be found in this directory (https://osf.io/3embr/). Figure.2.1 shows representative frame from all 4 activities.

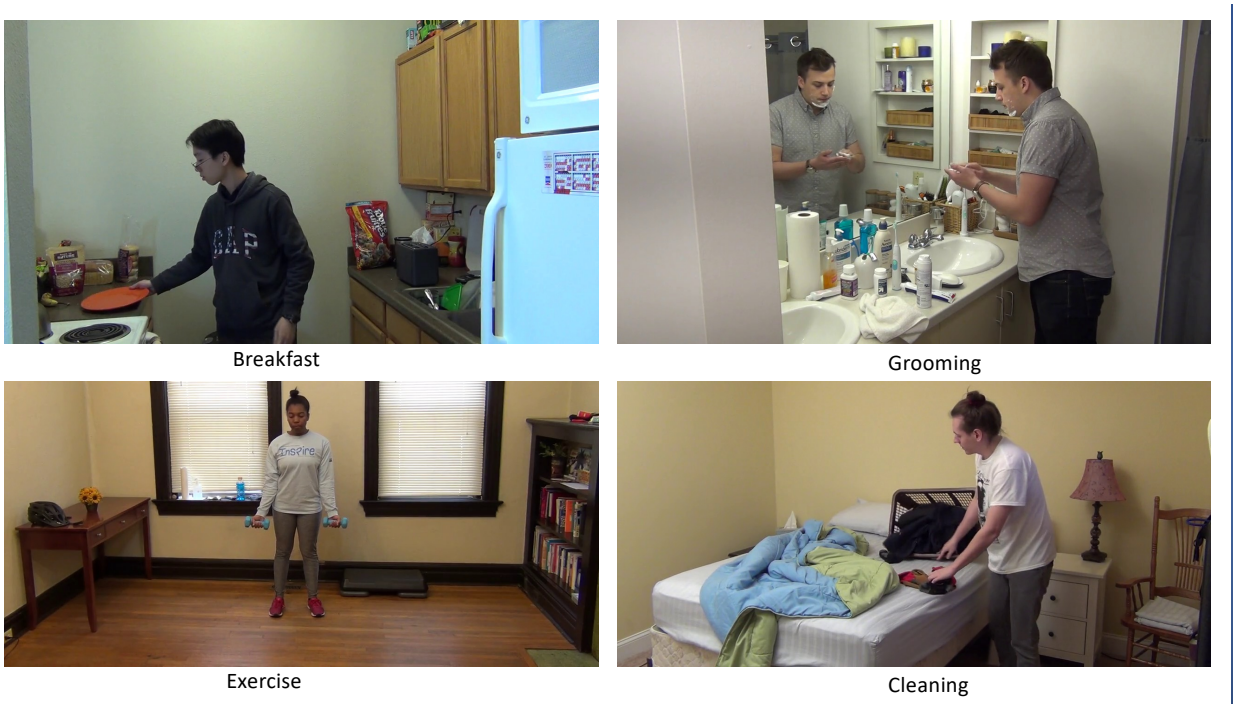

Figure 2.1: Representative Frames from the four movies..

2.1.3 Task and procedure

After giving informed consent to this study, participants passively watched four movies of an actor performing everyday activities. Breaks were offered between movies. Participants were instructed to watch the movies for comprehension while their eyes were recorded by the eye-tracker. The eye-tracker was calibrated by the experimenter before the start of the experiment, and during the break if needed. Gaze locations were obtained from the right eye using an infrared pupil-corneal eye tracker camera (EyeLink 1000; SR Research Ltd., Mississauga, ON, Canada) that sampled at 1000 Hz. The camera was mounted on the SR Research Desktop Mount. A chin/forehead rest was used to minimize head motion during the tasks. The camera was positioned 52 cm from the top of the forehead rest. Stimuli were presented on a 19-in. (74 cm) monitor (1440X3900 resolution, viewing distance of 58 cm from the forehead rest, viewing angle of 38.68).

Calibration of the eye tracker was conducted before beginning the study task. Participants were instructed to look at each of five to nine dots presented serially across the participant's central

and peripheral visual field. Following calibration, the measurements were validated by having the participants look at each of these nine dots again as they appeared on the screen. This validation of calibration was considered good when there was an average error of 0.50 degrees of visual angle or less and when the maximum error for any given dot was 1.00 degree or less.

Movie order was counterbalanced across participants. Participants were debriefed after completing the study.

2.2 Event Segmentation Norms

2.2.1 Participants

For the event segmentation part of the study, to make the distribution of segmentation probabilities less skewed, two separated datasets were combined to increasing the number of subjects. The first dataset included 184 participants from Amazon Mechanical Turk. Participants self-reported their age (mean age = 35 years, SD = 11.95 years), gender (70 female, 112 male, and 2 other). The recruitment procedure is detailed in another study. [27].

For the second dataset, participants were recruited from the Volunteer for Health participant registry, which is maintained by the Washington University School of Medicine. Data from 47 participants were included in the analyses for this study (age range: 18–35; mean age = 23 years; 14 males,33 female). Both studies received approval from the Washington University Human Research Protection Office and were conducted in accordance with the Declaration of Helsinki.

2.2.2 Materials

The same four movies from the passive viewing part of the study were used for this task.

2.2.3 Procedure

After giving informed consent, participants were randomly assigned either to identify coarse event boundaries or fine event boundaries in those four movies. In the coarse condition, participants were instructed to push the button whenever they believe that a large meaningful unit of activity has ended. In the fine condition, participants were instructed to push the button whenever they believe that a small meaningful unit of activity has ended.

All participants practiced the segmentation task before segmenting the four movies. During the practice session, participants segmented a video with a duration of 2 min 35 s in which a man constructs a toy boat using interlocking building blocks. Participants repeated the practice until they pressed the button within a pre-defined range: 5-8 times for fine segmentation and 2–4 times for coarse segmentation. Participants were never informed of these ranges but were asked to repeat the practice to identify more (or fewer) activities in the video until performance was within range.

2.3 Estimating Predictive Looking

To establish a continuous measure of predictive looking error, this study employed regression models to formalize predictive looking. The residuals of the predictive looking model served as the quantifiable prediction error. Predictive looking was defined as the extent to which participants' prior gaze locations predicted the current hand location of the actor. The hands' location was selected as the region of interest to approximate the most crucial area to predict in each frame, given the frequent hand movements observed across all movies. Attending to hand locations was deemed vital for processing and comprehending the four activities in this study. Moreover, previous research has indicated that individuals naturally focus on hand locations during naturalistic viewing, and distinctive hand positions are associated with event boundaries.

To facilitate model convergence and address the zero-inflation issue within the dataset, the model fitting procedure was conducted at the grid level rather than the pixel level. In essence, the

predictive looking model aimed to estimate the presence of the actor's hands within a grid square by leveraging the prior gaze positions of participants within that specific square. The creation of grids involved employing a set of vertical and horizontal lines to partition the frame into equal sections, with the width of each frame (1280 pixels) divided into 8 parts and the length of the frame (960 pixels) divided into 6 parts. Consequently, for each frame in every movie, this process resulted in the generation of 48 squares of equal sizes.

The dependent variable in the predictive looking model, denoted as H in equations 1-6, signifies the presence or absence of the actor's hands. To determine the existence of actor's hands for each grid in each frame for each movie, we initially identified the hand positions using the OpenPose algorithm, a computer vision model for pose detection [28]. Subsequently, the hand locations were convolved with a two-dimensional Gaussian kernel (sd=50 pixels, determined through visual inspection of the hands' density map to ensure the heatmap accurately represented hand locations). Following convolution, the sum of hand location probabilities for pixels within each grid was computed. An 'elbow point' in the distribution of these probabilities across all grids for each frame, representing a significant change in the rate of loss, was then identified. This elbow point served as the cut-off threshold. Grids surpassing this threshold were assigned a binary code of 1, indicating the presence of hands, while grids falling below the threshold were coded as zero.

The independent variables in the predictive looking model are represented by participants' gaze density for each grid in each frame, denoted as F in equation 1-6. To calculate gaze density for each grid in every frame, we initially convolved each participant's gaze location for a specific frame using a two-dimensional Gaussian kernel. The standard deviation of this Gaussian kernel was determined by the viewing angle of the participants (sd = 37 pixels). This convolution process generated the gaze density value for each pixel corresponding to each participant. Subsequently, the density values of all participants for all pixels within a given grid were aggregated, resulting in the final gaze density measurement for each grid.

To ensure that the model accurately represents predictive looking, grids containing the actor's

face were deliberately excluded from our analysis. This exclusion is grounded in the observation that, in movies, the position of the actor's face tends to exhibit relatively static behavior compared to hand movements, suggesting that gazing towards the face may not necessarily be indicative of predictive behavior. The same OpenPose algorithm [28] determined the location of the face. A two-dimensional gaussian kernel was applied onto the fact location. Following this, we identified an 'elbow point' within the distribution of these facial location probabilities across all grids. This 'elbow point' represents a significant change in the rate of loss, hence serving as our cut-off threshold. Any grids with values exceeding this threshold were subsequently removed from the analysis.

The same OpenPose algorithm [28] used for determining hand locations was employed to identify the location of the face. A two-dimensional Gaussian kernel was applied to the facial location, and subsequently, we pinpointed an 'elbow point' within the distribution of these facial location probabilities across all grids. This 'elbow point' signifies a substantial change in the rate of loss, thereby serving as our cut-off threshold. Any grids surpassing this threshold were consequently excluded from the analysis.

Equation 2.1 describes the form the predictive looking model:

$$H_{tr} = \beta_{i0}F_{t-ir} + \beta_{i-1,0}F_{t-(i+1)r} + \ldots + \beta_{00}F_{tr} + \epsilon \tag{2.1}$$

Where Htr is the hand location for region r (pixel) at the time point t(s); F represents the density of gaze locations. i indicates the furthest point in time where fixation probability predicts current hand location.

To enhance the model's robustness against potential random noise or perturbations and to facilitate easier interpretation, we implemented a series of steps to smooth out the temporal fluctuations in gaze and hand values. First, a Gaussian kernel was applied to smooth the probabilities associated with gaze and hand movements across the temporal span of each movie. The size of the smoothing kernel (120 frames) was determined through visual inspection. Subsequently, we sampled these smoothed values at intervals of every 60 frames. This sampling rate was chosen to align with the

frames per second (fps) of the movies and adhered to the standard temporal resolution commonly employed in dynamic cognition studies, which utilizes one-second intervals.

Given the binary nature of the dependent variable, we opted to employ multivariate mixed-effects logistic models in lme4 package (Version 1.1.27.1; Bates et al., 2015) in R software (R Core Team, 2014). The independent variables were the recent gaze densities for each grid, with the presence of the hand within that specific grid serving as the dependent variable. To account for potential variance stemming from specific movies and the diverse locations of the grid within frames, both grid index and movie were integrated as random intercepts within the models. To evaluate the degree of predictive looking, a stepwise model comparison was employed. A more sophisticated model incorporated the gaze density of the specific grid one second further into the past, while controlling for gaze densities closer to the current hand locations. Gaze densities from the past were sampled at a rate of one second. The stepwise model comparison approach can be conceptualized as follows:

$$H_{tr} = \beta_{00}F_{tr} + \epsilon \quad \text{(Equation 2)}$$

$$H_{tr} = \beta_{10}F_{t-1r} + \beta_{00}F_{tr} + \epsilon \quad \text{(Equation 3)}$$

$$H_{tr} = \beta_{20}F_{t-2r} + \beta_{10}F_{t-1r} + \beta_{00}F_{tr} + \epsilon \quad \text{(Equation 4)}$$

$$H_{tr} = \beta_{30}F_{t-3r} + \beta_{20}F_{t-2r} + \beta_{10}F_{t-1r} + \beta_{00}F_{tr} + \epsilon \quad \text{(Equation 5)}$$

$$\vdots$$

$$H_{tr} = \beta_{i0}F_{t-ir} + \beta_{i-1,0}F_{t-(i+1)r} + \ldots + \beta_{30}F_{t-3r} + \beta_{20}F_{t-2r} + \beta_{10}F_{t-1r} + \beta_{00}F_{tr} + \epsilon \quad \text{(Equation 6)}$$

Where Htr is the hand location for region r (in pixel) at the time point t (in seconds); F represents the density of gaze locations. i indicates the furthest point in time where fixation probability predicts the current hand location.

The optimal model was determined based on the relative variation in both Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values among these models. Prediction

errors were quantified as the magnitude of the deviance residuals for each grid in each frame based on the optimal predictive looking model. Deviance residuals in logistic regression are a measure of the difference between observed and expected outcomes under the model. The mathematical formula for the magnitude of the deviance residual is:

$$\text{Prediction Error} = \sqrt{-2\log(\text{Prob\_event})} \tag{2.2}$$

2.4 Predictive Signals of Computational Model of Event Comprehension

Two continuous computationally derived predictive signals, namely prediction error and prediction uncertainty, were extracted from SEM-PE and SEM-UNCERTAINTY, respectively. For detailed infrastructure specifications of SEM-PE and SEM-UNCERTAINTY, refer to [2]. The fundamental architecture of SEM-PE and SEM-UNCERTAINTY is identical. In both models, event schemas, representing knowledge of event types, are represented as RNN's weights matrices, and each event model (an internal representation of the current situation) is represented as RNN's matrices and hidden unit activations. In both models, an approximate Bayesian inference process assigns incoming scene vectors to different event schemas. On each time step, a currently active RNN is presented with the current scene vector and predicts the subsequent scene vector. An event boundary is operationalized as the transition from one RNN to another.

One critical distinction between SEM-PE and SEM-UNCERTAINTY lies in the gating signals for RNN (event schema) switching. In SEM-PE, prediction error functions as gating signal to determine when to evaluate alternative schemas. Prediction error was operationalized as Euclidean distance between an observed scene vector and the RNN's prediction. If prediction error surpasses a predefined threshold, the event schema inference process to determine alterative event model; otherwise, the current event schema is assigned to the current scene. Similarly, in the

SEM-UNCERTAINTY model, the switching process is initiated when prediction uncertainty exceeds threshold, the switching process is triggered. Prediction uncertainty was operationalized as variability in RNN predictions across perturbations of the RNN weights. The variance of these predicted scene vectors approximates prediction uncertainty induced by uncertainty about RNN's weights.

Chapter 3: Results

Data were analyzed to address five main questions 1) Can a continuous measure of prediction error be derived from gaze locations? To address this, we constructed statistical models that use past gaze locations to predict the current hand location of the actor. 2) What is the extent of participants' predictive looking? A step-wise model comparison approach was implemented to identify the most optimal predictive looking model to determine the extent of participants' predictive looking. 3) What is the relationship between predictive looking error and computational models of prediction error? To explore this question, we examined the correlation between the magnitude of the residual in the most optimal predictive looking model and the prediction error derived from the computational model (SEM-PE). Additionally, we also examined the relationship between the same magnitude of the residual with prediction uncertainty generated by another computational model (SEM-UNCERTAINTY). 4) Do people exhibit high agreement on event boundaries? To answer this, we quantified people's segmentation behavior throughout movie watching. And finally, 5) Is predictive looking error associated with segmentation probability? To address this question, we analyzed the relationship between the magnitude of residual in the most optimal predictive looking error with segmentation probability.

3.1 Participants Exhibited Predictive Looking During Passive Movie Viewing

To quantify predictive looking, a forward selection stepwise model comparison approach was implemented. The analysis began with the simplest model (Equation 2), where the current gaze density predicts the current hand existence in a grid. A model that is one level more sophisticated

than the simplest model is using gaze density of the grid one second ago in addition to the current gaze density of that grid to predict the current hand existence of the same grid. Progressing to more sophisticated models, each level considered an additional second of gaze density from the past, culminating in a model incorporating gaze density up to 9 seconds ago to predict the current hand existence in the same grid (equation 6). All models included random intercepts of grid locations and movie types. To assess model fit, both AIC and BIC index were calculated for each model. These indexes revealed that a model that includes gaze density of the grid up to 9 seconds in the past is the best model to predict the current hand location. The comparative AIC and BIC values across models are shown in Figure 3. AIC values showed a substantial decrease in AIC values until the model including gaze density up to 9 seconds. BIC values decreased until the model with gaze density from 5 seconds in the past, reaching stability until the 9-second model, after which they increased. This suggests that, across four everyday activities, participants exhibited predictive looking behavior up to 9 seconds.

3.2 Predictive Looking Error Correlates with Computational Model-based Prediction Error

Now that we have identified the best fitted predictive looking model, we can establish a continuous measure of prediction error. We operationalized prediction error as the magnitude of the deviance residuals as it represents the discrepancies between prediction and the reality. We will refer to this kind of prediction error as "predictive looking error" in the following paragraph, as it was derived from participants' gaze patterns.

To explore the relationship between predictive looking errors and prediction error generated by the computational mode SEM-PE, we conducted a mixed-effect linear regression. In this regression, predictive looking error served as the dependent variable, while SEM-PE's prediction error acted as the independent variable. This model included random intercepts for the type of movies and the

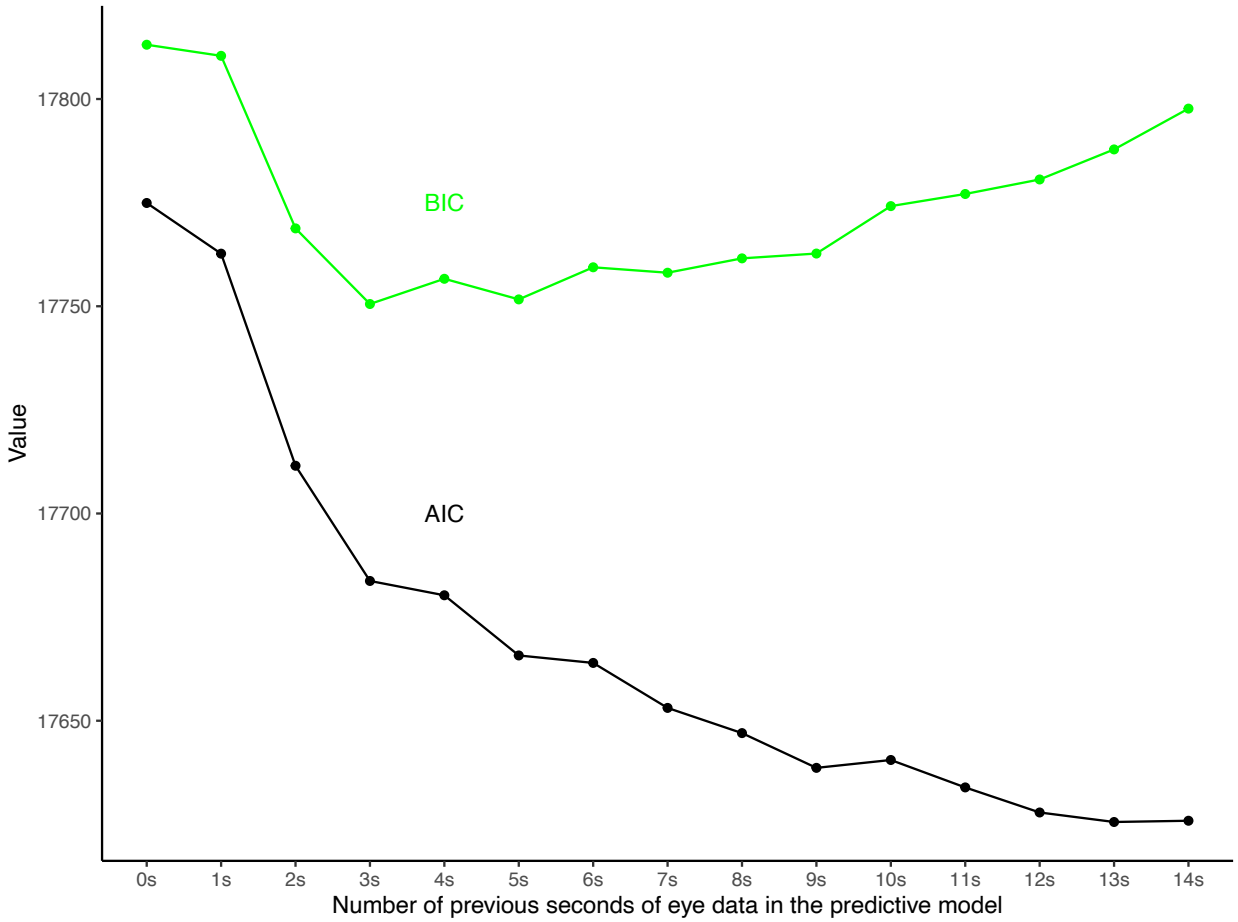

Figure 3.1: AIC and BIC values for predictive looking models include gaze densities up to the furthest time point in the past

grid number. The results revealed a significant and positive correlation between prediction error and predictive looking errors (prediction error model: β= 0.035, t-value= 13.62).

In addition, we explored the relationship between predictive looking error and prediction uncertainty to assess the specificity of this measure of prediction error. A similar mixed-effects linear regression was constructed, with predictive looking error as the dependent variable and the prediction uncertainty of SEM-UNCERTAINTY as the independent variable, controlling for the random effects of movie types and grid numbers. The results indicated a significant and positive correlation between prediction uncertainty and predictive looking errors (prediction error model: β= 0.028, t-value: 10.53). Figure 4 visually represents the relationship between eye gaze prediction

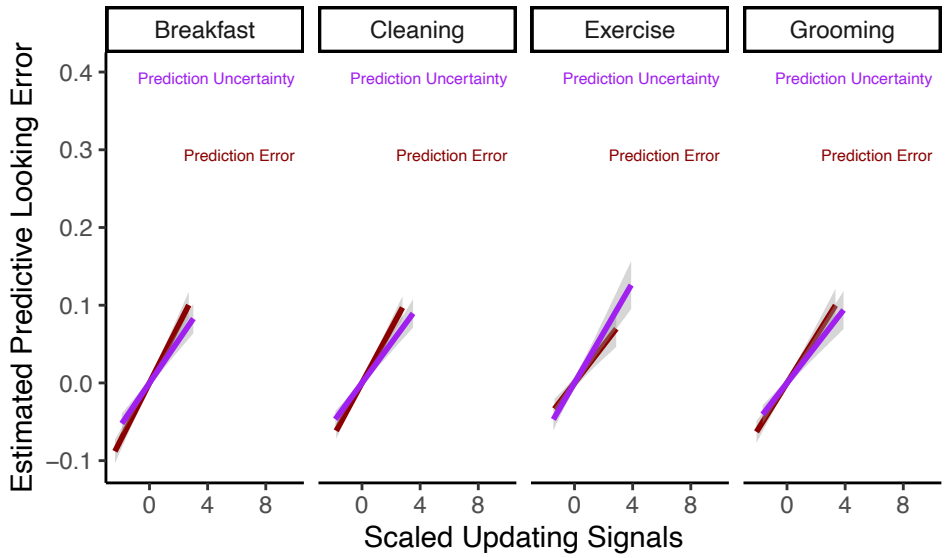

Figure 3.2: Linear regression lines are fitted between model estimated scaled predictive looking error and scaled updating signals (prediction error or prediction uncertainty) for all four movies. Color indicates the type of updating signal.

error values, prediction errors by SEM-PE, and prediction uncertainty by SEM-UNCERTAINTY for each movie. By comparing the two mixed-effects linear regressions (SEM-PE's prediction error predicting predictive looking error versus SEM-UNCERTAINTY's prediction uncertainty predicting predictive looking error), the AIC and BIC values suggest that the prediction error model is a better fit for predictive looking errors compared to uncertainty (prediction error model AIC: 267363.9, BIC: 267438.4; prediction uncertainty model AIC: 267438.4, BIC: 267486.2). According to Akaike weights, the model incorporating SEM-PE's prediction error is significantly superior to the one involving SEM-UNCERTAINTY's prediction uncertainty (Prediction error model's Akaike weight is 1, whereas the prediction uncertainty model's weight is 0).

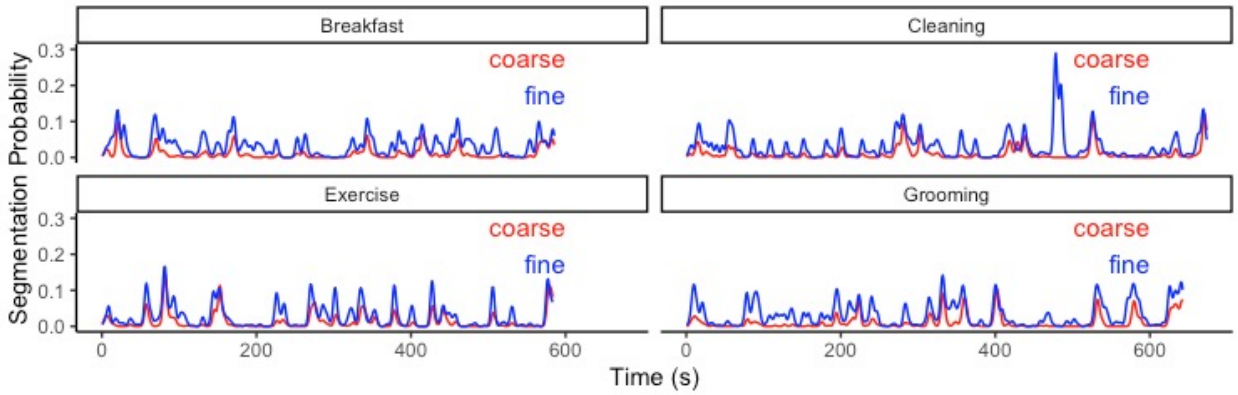

Figure 3.3: Segmentation Probabilities throughout four different movies for both fine and coarse conditions.

3.3 Predictive Looking Error is Associated with Segmentation Probability

To characterize participants' segmentation behavior, segmentation probabilities were calculated throughout each movie for both the fine and coarse conditions (Figure 2). To derive these probabilities, a Gaussian kernel was initially applied to all the time points when participants pressed the spacebar, generating segmentation probabilities for both conditions. As depicted in Figure 2, despite the intentionally vague instructions, certain time points within the movies were consistently identified by many participants as event boundaries, while at other times, no spacebar presses occurred. This observed pattern aligned with segmentation behavior reported in prior research [15, 26]

A mixed effects multivariate regression model was employed to assess the relationship between predictive looking error and segmentation probability where the grains of segmentation are entered as covariate , and the type of movies as well as grid numbers were entered as the random effects. The results was consistent with what the theory predicted and revealed a significant and positive prediction of segmentation probabilities by predictive looking errors (β= 0.001, t-value= 6.63). As expected, segmentation probabilities were significantly higher in the fine segmentation condition. Additionally, there was no significant interaction between conditions and predictive looking errors.

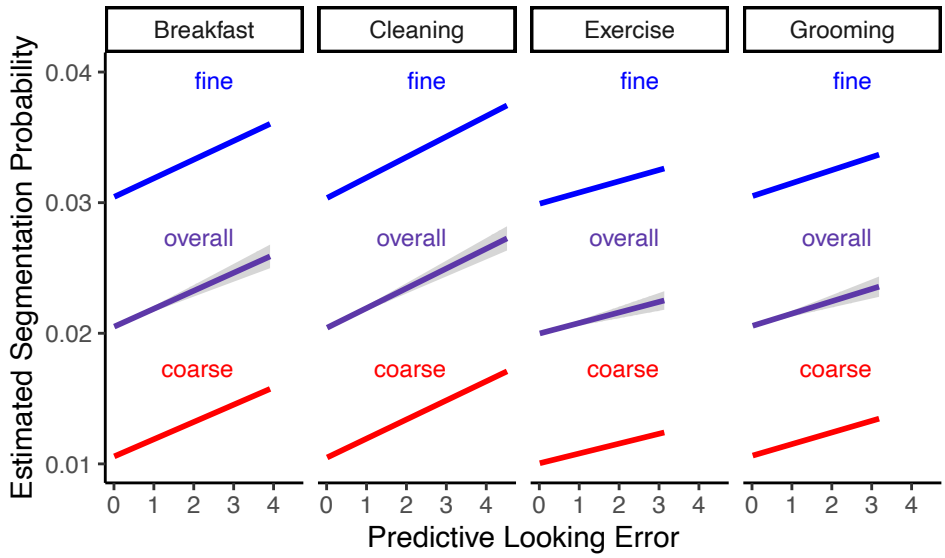

Figure 3.4: Linear regression lines are fitted between estimated looking error and segmentation probability for all four movies where color indicates the segmentation condition.

Figure 5 illustrates the relationship between predictive looking error and segmentation probability for each individual movie.

Chapter 4: Discussion

The present study uncovered five critical phenomena related to event perception: firstly, people's gaze position can unveil prediction error. Secondly, individuals exhibit predictive looking up to 9 seconds into the future while passively observing everyday activities. Thirdly, there was a significant and positive correlation between predictive looking errors and computationally generated prediction errors, indicating that both measures encapsulate the dynamics of prediction error in the brain. Fourthly, untrained participants demonstrated high agreement of event boundaries. Finally, predictive looking errors demonstrated an increase around event boundaries, providing support for the Event Segmentation Theory.

4.1 People Looked Predictively up to 9 Seconds During the Unfolding of Everyday Activities

Model selection revealed that eye gaze signals from as far back as nine seconds significantly improved the prediction of hand position. This indicates that viewers' brains engage in predictive processes extending up to nine seconds into the future during passive observation of daily activities. For instance, in the breakfast movie, participants started directing their gaze towards the toaster once the actor picked up the plates from the counter, anticipating the subsequent movement of the actor's hand in the upcoming seconds.

Prediction manifests in various forms of eye movements. For example, anticipatory pursuit can begin as early as 200 milliseconds before the onset of object motion when the motion direction and onset are predictable [29, 30]. Saccades, too, synchronize with target steps rather than exhibit a time lag if the target is predictable [31]. Beyond tracking single object or target in highly, individuals

demonstrate the ability to anticipate future events. In sports, professional cricket players fixate on the ball as it leaves the bowler's hand and then rapidly make a saccade to where the ball is expected to bounce, awaiting its image to return to the fovea [32]. In daily activities outside of sports, Eisenberg et al. [25] showed participants increased their gaze time on a target object as the actor approached, suggesting an anticipation of the actor making contact with the object in the near future. Building on previous research, the current data-driven approach unveils that predictive looking extends beyond critical objects or specific time intervals; rather, it is a continuous and integral aspect of natural perception.

4.2 Predictive Looking Errors Increase Around Event Boundaries

Previous studies have shown that event boundaries influence various characteristics of eye movements. Eisenberg et al. [25] documented that predictive looking towards critical objects near event boundaries tends to occur later and less frequently compared to looks directed towards objects within events. Clearly, event boundaries affect the timing and frequency of predictive looking. Moreover, viewers initiate an exploratory processing phase around event boundaries, before transitioning to focal viewing as the event progresses. This boundary-evoked ambient phase of processing is interpreted as adaptive and due to the unpredictability of activity around event boundaries[33].

Limitations of previous research was that predictively looking error was measured sparsely [25] or indirectly [33]. To overcome these limitations, our study introduced a direct and continuous method for estimating predictive looking errors during movie viewing. This novel measure was validated by showcasing superior model fit in a statistical model that integrated computational model-generated prediction error compared to prediction uncertainty. This outcome indicates that our measure of predictive looking errors effectively captures the true prediction errors unfolding in the brain.

Having derived and validated this measure of prediction error, we then showed that controlling for the fixed effect segmentation condition and random effect of movie types, there was a significant and positive relationship between predictive looking errors and segmentation probabilities. This finding provides strong evidence for theories of event cognition that suggest that prediction error drives event segmentation.

4.3 Implications for Theories of Event Perception

Currently, theories of event segmentation focus on three types of mechanisms for event segmentation: detecting prediction errors [1], detecting feature changes [34, 35] or detecting statistical structure [36, 37] Our analysis results indicate a positive correlation between an increase in prediction error and segmentation probabilities. We delve into the implications of this finding for each proposed mechanism below.

According to EST[1], observers continuously make predictions based on a working event model, monitoring discrepancies between the model's predictions and reality. When prediction errors increase, the model resets and updates to reflect the new situation, resulting in the perception of event boundaries between the old and new event models. Consistent with this theory, the current study demonstrates that individuals engage in predictive looking, and the errors in predictive looking are significantly and positively correlated with segmentation probabilities.

Other accounts of event segmentation emphasize learning the sequential structure to chunk continuous experience into units, such as predictable-unpredictability moments [38] and community structure [37]. For accounts that focus on predictable unpredictability e .g, [36], it would predict a time lag between event boundaries and the peak of prediction error, as an observer would anticipate the increase in prediction error in the imminent future and then segment before the peak of prediction error. The current study design is inadequate to test this alternative account because the lack of effect could be due to the response time required to execute the button press movement. For theories that focus on the community structure as a mechanism for segmentation, they would hypothesize

that people's predictive looking to stimuli is uncorrelated with segmentation probabilities if the transitional predictability between the stimuli's location is identical, making the prediction error consistent. However, in naturalistic videos, community structures are often confounded with fluctuation of prediction errors. Therefore, future experiments are needed to test this hypothesis.

Finally, this study's findings offer limited insights for accounts that propose event boundaries are determined through retroactive inferences [34, 35, 39] Baker and Levin [34] suggest the spatial configuration of a recently encountered scene is maintained in memory and compared to current perceptual input. Changes in spatial configurations lead to segmentation. Papenmeier et al [39] showed that participants segment fewer times when casual continuation information can be retroactively inferred. As this study focused on the predictive part of human perception and provided positive evidence for prediction error drives segmentation, future study is needed to explore whether subset of event boundaries could be due to retroactive change detection.

In summary, our study provided robust evidence that increase in prediction errors drive event segmentation, while future investigations are required to unravel whether a subset of event boundaries could due to detection of statistical structure or feature codes retrospectively.

This study has certain limitations that underscore the necessity for further investigation in this domain. First, the reliance on the location of the actor's hands as a proxy for the most important features or the target of prediction may not be universally applicable. This approach may not be applicable in scenarios where the actor's hands are not present or where their location is not important to the viewer. This limitation is evident in the exercise movie used in this study, where the hands' location is not the most informative feature throughout these activities, resulting in the observed absence of a relationship between prediction errors and segmentation probabilities.

One potential remedy to modeling saliency or the target of prediction is to leverage computational models to derive saliency maps for each frame and treat them as the target of prediction. However, a challenge lies in selecting an appropriate computational model for movie frames. Viewers' attention differs between static pictures and movies, as evidenced by previous studies showing that object

semantics play a critical role in guiding attention in static pictures [40]. On the other hand, when watching movies, people tend to focus on features that are dynamic, such as actors' hands. Therefore, using computational saliency models trained on static pictures may not adequately represent the target of prediction during movie viewing. On the other hand, it is plausible that computational saliency models specifically trained on movies may inherently capture predictive looking within the derived saliency maps.

References

1. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. Event Perception: A Mind-Brain Perspective. *Psychological Bulletin* **133,** 273–293. doi:10.1037/0033-2909.133.2.273 (2007) (cited on pp. 1, 2, 25).
2. Bezdek, M., Nguyen, T., Gershman, S. J., Bobick, A., Braver, T. S. & Zacks, J. M. *Error-Based Updating of Event Representations Enables Prediction of Human Activity at Human Scale* preprint (PsyArXiv, 2022). doi:10.31234/osf.io/pt6hx (cited on pp. 1, 4, 15).
3. Diaz, G., Cooper, J., Rothkopf, C. & Hayhoe, M. Saccades to Future Ball Location Reveal Memory-Based Prediction in a Virtual-Reality Interception Task. *Journal of Vision* **13,** 20–20. doi:10.1167/13.1.20 (2013) (cited on pp. 1, 5, 6).
4. Henderson, J. M. Gaze Control as Prediction. *Trends in Cognitive Sciences* **21,** 15–23. doi:10.1016/j.tics.2016.11.003 (2017) (cited on pp. 1, 6).
5. Itti, L. & Koch, C. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience* **2,** 194–203. doi:10.1038/35058500 (2001) (cited on p. 1).
6. Malcolm, G. L. Combining Top-down Processes to Guide Eye Movements during Real-World Scene Search. *Journal of Vision* **10,** 1–11. doi:10.1167/10.2.4 (2010) (cited on p. 1).
7. Mann, D. L., Spratford, W. & Abernethy, B. The Head Tracks and Gaze Predicts: How the World's Best Batters Hit a Ball. *PLoS ONE* **8** (ed Zeil, J.) e58289. doi:10.1371/journal.pone.0058289 (2013) (cited on p. 1).
8. Rothkopf, C. A., Ballard, D. H. & Hayhoe, M. M. Task and Context Determine Where You Look. *Journal of Vision* **7,** 16. doi:10.1167/7.14.16 (2016) (cited on p. 1).
9. Glimcher, P. W. Understanding Dopamine and Reinforcement Learning: The Dopamine Reward Prediction Error Hypothesis. *Proceedings of the National Academy of Sciences* **108,** 15647–15654. doi:10.1073/pnas.1014269108 (supplement_3 2011) (cited on p. 2).
10. Maia, T. V. Reinforcement Learning, Conditioning, and the Brain: Successes and Challenges. *Cognitive, Affective, & Behavioral Neuroscience* **9,** 343–364. doi:10.3758/CABN.9.4.343 (2009) (cited on p. 2).
11. Bubic. Prediction, Cognition and the Brain. *Frontiers in Human Neuroscience.* doi:10.3389/fnhum.2010.00025 (2010) (cited on p. 2).
12. Friston, K., Kilner, J. & Harrison, L. A Free Energy Principle for the Brain. *Journal of Physiology-Paris* **100,** 70–87. doi:10.1016/j.jphysparis.2006.10.001 (2006) (cited on p. 2).

13. Maldonato, M. & Dell'Orco, S. The Predictive Brain. *World Futures* **68,** 381–389. doi:10.1080/02604027.2012.693846 (2012) (cited on p. 2).

14. Niv, Y. & Schoenbaum, G. Dialogues on Prediction Errors. *Trends in Cognitive Sciences* **12,** 265–272. doi:10.1016/j.tics.2008.03.006 (2008) (cited on p. 2).

15. Zacks, J. M., Tversky, B. & Iyer, G. Perceiving, Remembering, and Communicating Structure in Events. *Journal of Experimental Psychology: General* **130,** 29–58. doi:10.1037/0096-3445.130.1.29 (2001) (cited on pp. 3, 21).

16. Newtson, D., Engquist, G. & Bois, J. The Objective Basis of Behavior Units (1973) (cited on p. 3).

17. Speer, N. K., Swallow, K. M. & Zacks, J. M. Activation of Human Motion Processing Areas during Event Perception. *Cognitive, Affective, & Behavioral Neuroscience* **3,** 335–345. doi:10.3758/CABN.3.4.335 (2003) (cited on p. 3).

18. Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U. & Norman, K. A. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95,** 709–721.e5. doi:10.1016/j.neuron.2017.06.041 (2017) (cited on p. 3).

19. Kuperberg, G. R. Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science* **13,** 256–298. doi:10.1111/tops.12518 (2021) (cited on p. 4).

20. Shin, Y. S. & DuBrow, S. Structuring Memory Through Inference□Based Event Segmentation. *Topics in Cognitive Science* **13,** 106–127. doi:10.1111/tops.12505 (2021) (cited on p. 4).

21. Jelinek, F. Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE* **64,** 532–556 (1976) (cited on p. 5).

22. Norris, D. & McQueen, J. M. Shortlist B: A Bayesian Model of Continuous Speech Recognition. *Psychological Review* **115,** 357–395. doi:10.1037/0033-295X.115.2.357 (2008) (cited on p. 5).

23. Kleinschmidt, D. F. & Jaeger, T. F. Robust Speech Perception: Recognize the Familiar, Generalize to the Similar, and Adapt to the Novel. *Psychological review* **122,** 148 (2015) (cited on p. 5).

24. Romberg, A. R. & Saffran, J. R. Expectancy Learning from Probabilistic Input by Infants. *Frontiers in Psychology* **3.** doi:10.3389/fpsyg.2012.00610 (2013) (cited on p. 5).

25. Eisenberg, M. L., Zacks, J. M. & Flores, S. Dynamic Prediction during Perception of Everyday Events. *Cognitive Research: Principles and Implications* **3,** 53. doi:10.1186/s41235-018-0146-z (2018) (cited on pp. 6, 7, 24).

26. Gold, D. A., Zacks, J. M. & Flores, S. Effects of Cues to Event Segmentation on Subsequent Memory. *Cognitive Research: Principles and Implications* **2,** 1. doi:10.1186/s41235-016-0043-2 (2017) (cited on pp. 8, 21).

27. Bezdek, M., Nguyen, T. T., Hall, C. S., Braver, T. S., Bobick, A. F. & Zacks, J. M. The Multi-Angle Extended Three-Dimensional Activities (META) Stimulus Set: A Tool for Studying Event Cognition. *Behavior Research Methods.* doi:10.3758/s13428-022-01980-8 (2022) (cited on p. 10).

28. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. & Sheikh, Y. A. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cited on pp. 12, 13).

29. Badler, J. B. & Heinen, S. J. Anticipatory Movement Timing Using Prediction and External Cues. *The Journal of Neuroscience* **26,** 4519–4525. doi:10.1523/JNEUROSCI.3739-05.2006 (2006) (cited on p. 23).

30. Barnes, G. R. & Collins, C. Evidence for a Link Between the Extra-Retinal Component of Random-Onset Pursuit and the Anticipatory Pursuit of Predictable Object Motion. *Journal of Neurophysiology* **100,** 1135–1146. doi:10.1152/jn.00060.2008 (2008) (cited on p. 23).

31. Smit, A. C. & Van Gisbergen, J. A. M. A Short-Latency Transition in Saccade Dynamics during Square-Wave Tracking and Its Significance for the Differentiation of Visually-Guided and Predictive Saccades. *Experimental Brain Research* **76,** 64–74 (1989) (cited on p. 23).

32. Land, M. F. & McLeod, P. From Eye Movements to Actions: How Batsmen Hit the Ball. *Nature Neuroscience* **3,** 1340–1345. doi:10.1038/81887 (2000) (cited on p. 24).

33. Eisenberg, M. L. & Zacks, J. M. Ambient and Focal Visual Processing of Naturalistic Activity. *Journal of Vision* **16,** 5. doi:10.1167/16.2.5 (2016) (cited on p. 24).

34. Baker, L. J. & Levin, D. T. The Role of Relational Triggers in Event Perception. *Cognition* **136,** 14–29. doi:10.1016/j.cognition.2014.11.030 (2015) (cited on pp. 25, 26).

35. Hymel, A., Levin, D. T. & Baker, L. J. Default Processing of Event Sequences. *Journal of Experimental Psychology: Human Perception and Performance* **42,** 235–246. doi:10.1037/xhp0000082 (2016) (cited on pp. 25, 26).

36. Baldwin, D., Andersson, A., Saffran, J. & Meyer, M. Segmenting Dynamic Human Action via Statistical Structure. *Cognition* **106,** 1382–1407. doi:10.1016/j.cognition.2007.07.005 (2008) (cited on p. 25).

37. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural Representations of Events Arise from Temporal Community Structure. *Nature Neuroscience* **16,** 486–492. doi:10.1038/nn.3331 (2013) (cited on p. 25).

38. Baldwin, D. & Kosie, J. E. How Does the Mind Render Streaming Experience as Events? *Topics in Cognitive Science* **13,** 79–105. doi:10.1111/tops.12502 (2021) (cited on p. 25).

39. Papenmeier, F., Brockhoff, A. & Huff, M. Filling the Gap despite Full Attention: The Role of Fast Backward Inferences for Event Completion. *Cognitive Research: Principles and Implications* **4,** 3. doi:10.1186/s41235-018-0151-2 (2019) (cited on p. 26).

40. Hayes, T. R. & Henderson, J. M. Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes (2021) (cited on p. 27).