

1 Convolutional neural networks can decode eye movement data: A black box approach to
2 predicting task from eye movements

³ Zachary J. Cole¹, Karl M. Kuntzman¹, Michael D. Dodd¹, & Matthew R. Johnson¹

⁴ ¹ University of Nebraska-Lincoln

Author Note

The data used for the exploratory and confirmatory analyses in the present manuscript
are derived from experiments funded by NIH/NEI Grant 1R01EY022974 to MDD. Work
done to develop the analysis approach was supported by NSF/EPSCoR grant #1632849
(MRJ and MDD). Additionally, this work was supported by the National Institute of General
Medical Sciences of the National Institutes of Health [grant number P20 GM130461 awarded
to MRJ and colleagues] and the Rural Drug Addiction Research Center at the University of
Nebraska-Lincoln. The content is solely the responsibility of the authors and does not
necessarily represent the official views of the National Institutes of Health or the University
of Nebraska.

Correspondence concerning this article should be addressed to Zachary J. Cole, 238
Burnett Hall, Lincoln, NE 68588-0308. E-mail: zachary@neurophysicole.com

17

Abstract

18 Previous attempts to classify task from eye movement data have relied on model
19 architectures designed to emulate theoretically defined cognitive processes, and/or data that
20 has been processed into aggregate (e.g., fixations, saccades) or statistical (e.g., fixation
21 density) features. *Black box* convolutional neural networks (CNNs) are capable of identifying
22 relevant features in raw and minimally processed data and images, but difficulty interpreting
23 the mechanisms underlying these model architectures has contributed to challenges in
24 generalizing lab-trained CNNs to applied contexts. In the current study, a CNN classifier
25 was used to classify task from two eye movement datasets (Exploratory and Confirmatory)
26 in which participants searched, memorized, or rated indoor and outdoor scene images. The
27 Exploratory dataset was used to tune the hyperparameters of the model, and the resulting
28 model architecture was re-trained, validated, and tested on the Confirmatory dataset. The
29 data were formatted into raw timeline data (i.e., x-coordinate, y-coordinate, pupil size) and
30 minimally processed images. To further understand the relative informational value of the
31 raw components of the eye movement data, the timeline and image datasets were broken
32 down into subsets with one or more of the components of the data systematically removed.
33 Average classification accuracies were compared between datasets and subsets. Classification
34 of the timeline data consistently outperformed the image data. The Memorize condition was
35 most often confused with the Search and Rate conditions. Pupil size was the least uniquely
36 informative eye movement component when compared with the x- and y-coordinates. The
37 general pattern of results for the Exploratory dataset was replicated in the Confirmatory
38 dataset. Overall, the present study provides a practical and reliable black box solution to
39 classifying task from eye movement data.

40

Keywords: deep learning, eye tracking, convolutional neural network, cognitive state,
41 endogenous attention

42

Introduction

43 The association between eye movements and mental activity is a fundamental topic of
44 interest in attention research that has provided a foundation for developing a wide range of
45 human assistive technologies. Early work by Yarbus (1967) showed that eye movement
46 patterns appear to differ qualitatively depending on the task-at-hand (for a review of this
47 work, see Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010). A replication of this work by
48 DeAngelus and Pelz (2009) showed that the differences in eye movements between tasks can
49 be quantified, and appear to be somewhat generalizable. Technological advances and
50 improvements in computing power have allowed researchers to make inferences regarding the
51 mental state underlying eye movement data, also known as the “inverse Yarbus process”
52 (Haji-Abolhassani & Clark, 2014).

53 Current state-of-the-art machine learning and neural network algorithms are capable of
54 identifying diagnostic patterns for the purpose of decoding a variety of data types, but the
55 inner workings of the resulting model solutions are difficult or impossible to interpret.
56 Algorithms that provide such solutions are referred to as *black box* models. Dissections of
57 black box models have been largely uninformative (Zhou, Bau, Oliva, & Torralba, 2019),
58 limiting the potential for researchers to apply the mechanisms underlying successful
59 classification of the data. Still, black box models provide a powerful solution for
60 technological applications such as human-computer interfaces (HCI; for a review, see
61 Lukander, Toivanen, & Puolamäki, 2017). While the internal operations of the model
62 solutions used for HCI applications do not necessarily need to be interpretable to serve their
63 purpose, Lukander et al. (2017) pointed out that the inability to interpret the mechanisms
64 underlying the function of black box solutions impedes the generalizability of these methods,
65 and increases the difficulty of expanding these findings to real life applications. To ground
66 these solutions, researchers guide decoding efforts by using eye movement data and/or
67 models with built-in theoretical assumptions. For instance, eye movement data is processed

68 into meaningful aggregate properties such as fixations or saccades, or statistical features such
69 as fixation density, and the models used to decode these data are structured based on the
70 current understanding of relevant cognitive or neurobiological processes (e.g., MacInnes,
71 Hunt, Clarke, & Dodd, 2018). Despite the proposed disadvantages of black box approaches
72 to classifying eye movement data, there is no clear evidence to support the notion that the
73 grounded solutions described above are actually more valid or definitive than a black box
74 solution.

75 The scope of theoretically informed solutions to decoding eye movement data is limited
76 to the extent of the current theoretical knowledge linking eye movements to cognitive and
77 neurobiological processes. As our theoretical understanding of these processes develops, older
78 theoretically informed models become outdated. Furthermore, these solutions are susceptible
79 to any inaccurate preconceptions that are built into the theory. Consider the case of Greene,
80 Liu, and Wolfe (2012), who were not able to classify task from commonly used aggregate eye
81 movement features (i.e., number of fixations, mean fixation duration, mean saccade
82 amplitude, percent of image covered by fixations) using correlations, a linear discriminant
83 model, and a support vector machine (see Table 1). This led Greene and colleagues to
84 question the robustness of Yarbus's (1967) findings, inspiring a slew of responses that
85 successfully decoded the same dataset by aggregating the eye movements into different
86 feature sets or implementing different model architectures (see Table 1; Haji-Abolhassani &
87 Clark, 2014; Borji & Itti, 2014; Kanan, Ray, Bseiso, Hsiao, & Cottrell, 2014). The
88 subsequent re-analyses of these data support Yarbus (1967) and the notion that mental state
89 can be decoded from eye movement data using a variety of combinations of data features and
90 model architectures. Collectively, these re-analyses did not point to an obvious global
91 solution capable of clarifying future approaches to the inverse Yarbus problem beyond what
92 could be inferred from black box model solutions, but did provide a wide-ranging survey of a
93 variety of methodological features that can be applied to theoretical or black box approaches.

Eye movements can only delineate tasks to the extent that the cognitive processes underlying the tasks can be differentiated (Król & Król, 2018). Every task is associated with a unique set of cognitive processes (Coco & Keller, 2014; Król & Król, 2018), but in some cases, the cognitive processes for different tasks may produce indistinguishable eye movement patterns. To differentiate the cognitive processes underlying task-evoked eye movements, some studies have chosen to classify tasks that rely on stimuli that prompt easily distinguishable eye movements, such as reading text (e.g., Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013). The eye movements elicited by salient stimulus features facilitate task classifications; however, because these eye movements are the consequence of a feature (or features) inherent to the stimulus rather than the task, it is unclear if these classifications are attributable to the stimulus or a complex mental state (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016). Additionally, the distinct nature of exogenously elicited eye movements prompts decoding algorithms to prioritize these bottom-up patterns in the data over higher-level top-down effects (Borji & Itti, 2014). This means that these models are identifying the type of information that is being processed, but are not necessarily reflecting the mental state of the individual observing the stimulus. Eye movements that are the product of bottom-up processes have been reliably decoded, which is relevant for some HCI applications; however, such efforts do not fit the spirit of the inverse Yarbus problem, which is concerned with decoding high-level abstract mental operations that are not dependent on particular stimuli.

Currently, there is not a clearly established upper limit to how well cognitive task can be classified from eye movement data. Prior evidence has shown that the task-at-hand is capable of producing distinguishable eye movement features such as the total scan path length, total number of fixations, and the amount of time to the first saccade (Castelhano, Mack, & Henderson, 2009; DeAngelus & Pelz, 2009). Decoding accuracies within the context of determining task from eye movements typically range from chance performance to relatively robust classification (see Table 1). In one case, Coco and Keller (2014) categorized

the same eye movement features used by Greene et al. (2012) with respect to the relative contribution of latent visual or linguistic components of three tasks (visual search, name the picture, name objects in the picture) with 84% accuracy (chance = 33%). While this manipulation is reminiscent of other experiments relying on the bottom-up influence of words and pictures (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016) the eye movements in the Coco and Keller (2014) tasks can be attributed to the occurrence of top-down attentional processes. A conceptually related follow-up to this study classified tasks along two spatial and semantic dimensions, resulting in 51% classification accuracy (chance = 25%; Król & Król, 2018). A closer look at these results showed that the categories within the semantic dimension were consistently misclassified, suggesting that this level of distinction may require a richer dataset, or a more powerful decoding algorithm. Altogether, there is no measurable index of relative top-down or bottom-up influence, but this body of literature suggests that the relative influence of top-down and bottom-up attentional processes may have a role in determining the decodability of the eye movement data.

As shown in Table 1, when eye movement data are prepared for classification, fixation and saccade statistics are typically aggregated along spatial or temporal dimensions, resulting in variables such as fixation density or saccade amplitude (Castelhano et al., 2009; MacInnes et al., 2018; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). The implementation of these statistical methods is meant to explicitly provide the decoding algorithm with characteristics of the eye movement data that are representative of theoretically relevant cognitive processes. For example, MacInnes et al. (2018) attempted to provide an algorithm with data designed to be representative of inputs to the frontal eye fields. In some instances, such as the case of Król and Król (2018), grounding the data using theoretically driven aggregation methods may require sacrificing granularity in the dataset. This means that aggregating the data has the potential to wash out certain fine-grained distinctions that could otherwise be detected. Data structures of any kind can only be decoded to the extent to which the data are capable of representing differences between

Table 1

Previous Attempts to Classify Cognitive Task Using Eye Movement Data

Study	Tasks	Features	Model Architecture	Accuracy (Chance)
Greene et al. (2012)	memorize, decade, people, wealth	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, dwell times	linear discriminant, correlation, SVM	25.9% (25%)
Haji-Abolhassani & James (2014)	Greene et al. tasks	fixation clusters	Hidden Markov Models	59.64% (25%)
Kanan et al. (2014)	Greene et al. tasks	mean fixation durations, number of fixations	multi-fixation pattern analysis	37.9% (25%)
Borji & Itti (2014)	Greene et al. tasks	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, first five fixations, fixation density	kNN, RUSBoost	34.34% (25%)
Borji & Itti (2014)	Yarbus tasks (i.e., view, wealth, age, prior activity, clothes, location, time away)	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, first five fixations, fixation density	kNN, RUSBoost	24.21% (14.29%)
Coco & Keller (2014)	search, name picture, name object	Greene et al. features, latency of first fixation, first fixation duration, mean fixation duration, total gaze duration, initiation time, mean saliency at fixation, entropy of attentional landscape	MM, LASSO, SVM	84% (33%)
MacInnes et al. (2018)	view, memorize, search, rate	saccade latency, saccade duration, saccade amplitude, peak saccade velocity, absolute saccade angle, pupil size	augmented Naive Bayes Network	53.9% (25%)
Król & Król (2018)	people, indoors/outdoors, white/black, search	eccentricity, screen coverage	feed forward neural network	51.4% (25%)

¹⁴⁸ categories. Given that the cognitive processes underlying distinct tasks are often overlapping¹⁴⁹ (Coco & Keller, 2014), decreasing the granularity of the data may actually limit the potential

150 of the algorithm to make fine-grained distinctions between diagnostic components underlying
151 the tasks to be decoded.

152 The current state of the literature does not provide any firm guidelines for determining
153 what eye movement features are most meaningful, or what model architectures are best
154 suited for determining mental state from eye movements. The examples provided in Table 1
155 used a variety of eye movement features and model architectures, most of which were
156 effective to some extent. A proper comparison of these outcomes is difficult because these
157 datasets vary in levels of chance and data quality. Datasets with more tasks to be classified
158 have lower levels of chance, lowering the threshold for successful classification. Additionally,
159 datasets with a lower signal-to-noise ratio will have a lower achievable classification accuracy.
160 For these reasons, outside of re-analyzing the same datasets, there is no consensus on how to
161 establish direct comparisons of these model architectures. Given the inability to directly
162 compare the relative effectiveness of the various theoretical approaches present in the
163 literature, the current study addressed the inverse Yarbus problem by allowing a black box
164 model to self-determine the most informative features from minimally processed eye
165 movement data.

166 The current study explored pragmatic solutions to the problem of classifying task from
167 eye movement data by submitting unprocessed x-coordinate, y-coordinate, and pupil size
168 data to a convolutional neural network (CNN) model. Instead of transforming the data into
169 theoretically defined units, we allowed the network to learn meaningful patterns in the data
170 on its own. CNNs have a natural propensity to develop low-level feature detectors similar to
171 the primary visual cortex (e.g., Seeliger et al., 2018); for this reason, they are commonly
172 implemented for image classification. To test the possibility that the image data are better
173 suited to the CNN classifier, the data were also transformed from raw timelines into simple
174 image representations. To our knowledge, no study has attempted to address the inverse
175 Yarbus problem using any combination of the following methods: (1) Non-aggregated data,

176 (2) image data format, and (3) a black-box CNN architecture. Given that CNN architectures
177 are capable of learning features represented in raw data formats, and are well-suited to
178 decoding multidimensional data that have a distinct spatial or temporal structure, we
179 expected that a non-theoretically-constrained CNN architecture could be capable of decoding
180 data at levels consistent with the current state of the art. Furthermore, despite evidence that
181 black box approaches to the inverse Yarbus problem can impede generalizability (Lukander
182 et al., 2017), we expected that when testing the approach on an entirely separate dataset,
183 providing the model with minimally processed data and the flexibility to identify the unique
184 features within each dataset would result in the replication of our initial findings.

185

Method

186 **Participants**

187 Two separate datasets were used to develop and test the deep CNN architecture. The
188 two datasets were collected from two separate experiments, which we refer to as Exploratory
189 and Confirmatory. The participants for both datasets consisted of college students
190 (Exploratory $N = 124$; Confirmatory $N = 77$) from the University of Nebraska-Lincoln who
191 participated in exchange for class credit. Participants who took part in the Exploratory
192 experiment did not participate in the Confirmatory experiment. All materials and
193 procedures were approved by the University of Nebraska-Lincoln Institutional Review Board
194 prior to data collection.

195 **Materials and Procedures**

196 Each participant viewed a series of indoor and outdoor scene images while carrying out
197 a search, memorization, or rating task. For the memorization task, participants were
198 instructed to memorize the image in anticipation of a forced choice recognition test. At the
199 end of each Memorize trial, the participants were prompted to indicate which of two images
200 was just presented. The two images were identical outside of a small change in the display

201 (e.g. object removed or added to the scene). For the rating task, participants were asked to
202 think about how they would rate the image on a scale from 1 (very unpleasant) to 7 (very
203 pleasant). The participants were prompted to provide a rating immediately after viewing the
204 image. For the search task, participants were instructed to find a small “Z” or “N” embedded
205 in the image. In reality, targets were not present in the images outside of a small subset of
206 images ($n = 5$) that were not analyzed but were included in the experiment design so
207 participants believed a target was always present. Trials containing the target were excluded
208 because search behavior was likely to stop if the target was found, adding considerable noise
209 to the eye movement data. For consistency between trial types, participants were prompted
210 to indicate if they found a “Z” or “N” at the end of each Search trial.

211 The same materials were used in both experiments with a minor variation in the
212 procedures. In the Confirmatory experiment, participants were directed as to where search
213 targets might appear in the image (e.g., on flat surfaces). No such instructions were provided
214 in the Exploratory experiment.

215 In both experiments, participants completed one mixed block of 120 trials (task cued
216 prior to each trial), or three uniform blocks of 40 trials (task cued prior to each block for a
217 total of 120 trials). Block type was assigned in counterbalanced order. When the blocks were
218 mixed, the trial types were randomly intermixed within the block. For uniform blocks, each
219 block consisted entirely of one of the three conditions (Search, Memorize, Rate), with block
220 types presented in random order. Each stimulus image was presented for 8 seconds. The
221 pictures were presented in color, with a size of 1024 x 768 pixels, subtending a visual angle of
222 $23.8^\circ \times 18.0^\circ$.

223 Eye movements were recorded using an SR Research EyeLink 1000 eye tracker with a
224 sampling rate of 1000Hz. Only the right eye was recorded. The system was calibrated using
225 a nine-point accuracy and validity test. Errors greater than 1° or averaging greater than 0.5°
226 in total were re-calibrated.

227 **Datasets**

228 On some trials, a probe was presented on the screen six seconds after the onset of the
229 trial which required participants to fixate the probe once detected. To avoid confounds
230 resulting from the probe, only the first six seconds of the data for each trial was analyzed.
231 Trials that contained fewer than 6000 samples within the first six seconds of the trial were
232 excluded before analysis. For both datasets, the trials were pooled across participants. After
233 excluding trials, the Exploratory dataset consisted of 12,177 of the 16,740 total trials, and
234 the Confirmatory dataset consisted of 9,301 of the 10,395 total trials.

235 The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling
236 time point in the trial were used as inputs to the deep learning classifier. These data were
237 also used to develop plot image datasets that were classified separately from the raw timeline
238 datasets. For the plot image datasets, the timeline data for each trial were converted into
239 scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data
240 point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location
241 of the dot, pupil size was used to determine the relative size of the dot, and shading of the
242 dot was used to indicate the time-course of the eye movements throughout the trial. The
243 background of the plot images and first data point were white. Each subsequent data point
244 was one shade darker than the previous data point until the final data point was reached.
245 The final data point was black. For standardization, pupil size was divided by 10, and one
246 unit was added. The plots were sized to match the dimensions of the data collection monitor
247 (1024 x 768 pixels) and then shrunk to (240 x 180 pixels) in an effort to reduce the
248 dimensionality of the data.

249 **Data Subsets.** The full timeline dataset was structured into three columns
250 representing the x- and y- coordinates, and pupil size for each data point collected in the
251 first six seconds of each trial. To systematically assess the predictive value of each XYP (i.e.,
252 x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image

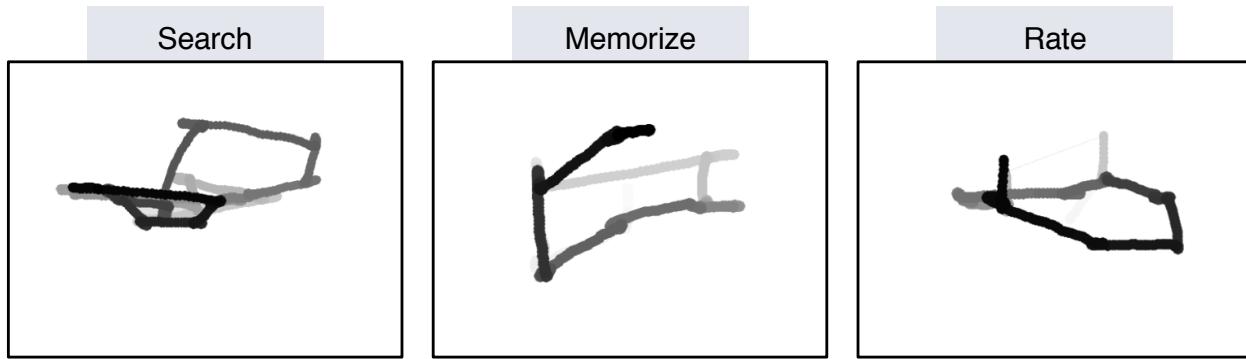


Figure 1. Each trial was represented as an image. Each sample collected within the trial was plotted as a dot in the image. Pupil size was represented by the size of the dot. The time course of the eye movements was represented by the gradual darkening of the dot over time.

253 datasets were batched into subsets that excluded one of the components (i.e., XYØ, XØP,
 254 ØYP), or contained only one of the components (i.e., XØØ, ØYØ, ØØP). For the timeline
 255 datasets, this means that the columns to be excluded in each data subset were replaced with
 256 zeros. The data were replaced with zeros because removing the columns would change the
 257 structure of the data. The same systematic batching process was carried out for the image
 258 dataset. See Figure 2 for an example of each of these image data subsets.

259 **Classification**

260 Deep CNN model architectures were implemented to classify the trials into Search,
 261 Memorize, or Rate categories. Because CNNs act as a digital filter sensitive to the number of
 262 features in the data, the differences in the structure of the timeline and image data formats
 263 necessitated separate CNN model architectures. The model architectures were developed
 264 with the intent of establishing a generalizable approach to classifying cognitive processes
 265 from eye movement data.

266 The development of these models was not guided by any formal theoretical
 267 assumptions regarding the patterns or features likely to be extracted by the classifier. Like
 268 many HCI models, the development of these models followed general intuitions concerned
 269 with building a model architecture capable of transforming the data inputs into an
 270 interpretable feature set that would not overfit the dataset. The models were developed

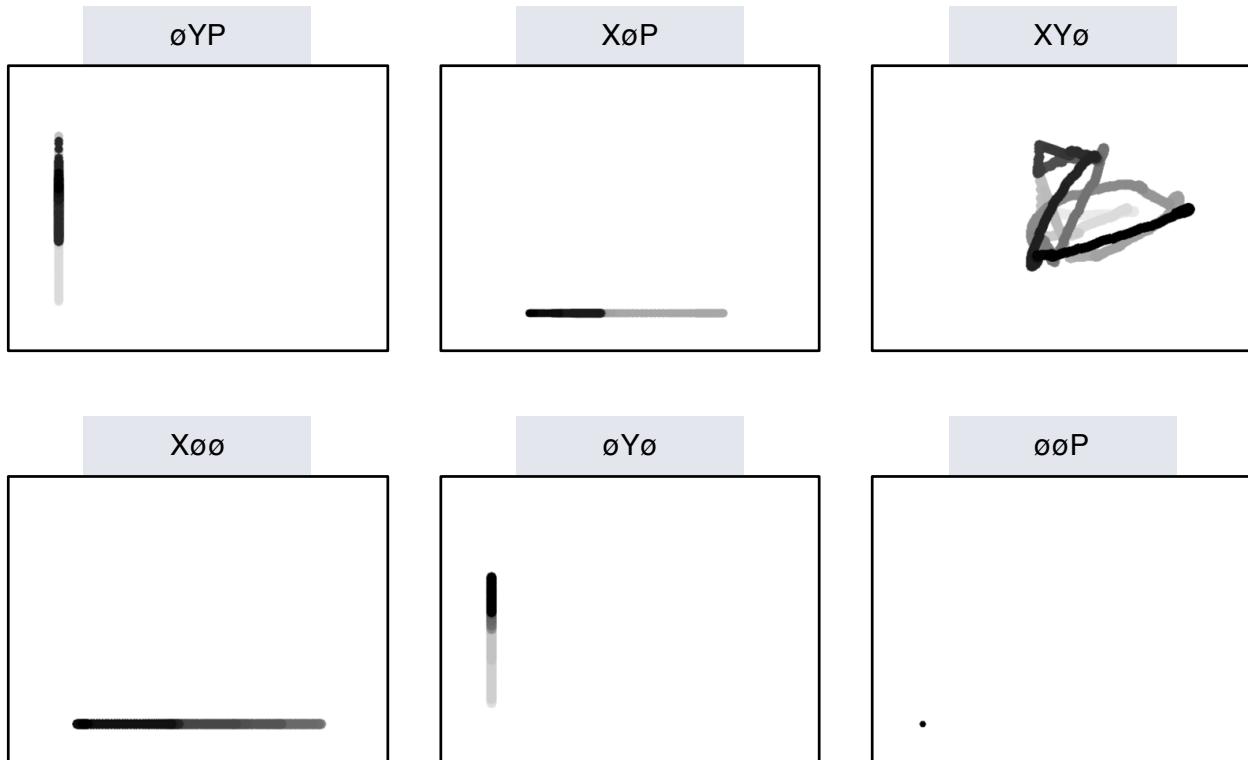


Figure 2. Plot images were used to represent data subsets that excluded one component of the eye movement data (i.e., XYØ, XØØ, ØYP) or contained only one component (i.e., XØØ, ØYØ, ØØP). As with the trials in the full XYP dataset, the time course of the eye movements was represented by the shading of the dot. The first sample of each trial was white, and the last sample was black.

271 using version 0.3b of the DeLINEATE toolbox, which operates over a Keras backend
 272 (<http://delineate.it>; Kuntzman et al., under review). Each training/test iteration randomly
 273 split the data so that 70% of the trials were allocated to training, 15% to validation, and
 274 15% to testing. Training of the model was stopped when validation accuracy did not improve
 275 over the span of 100 epochs. Once the early stopping threshold was reached, the resulting
 276 model was tested on the held-out test data. This process was repeated 10 times for each
 277 model, resulting in 10 classification accuracy scores for each model. The resulting accuracy
 278 scores were used for the comparisons against chance and other datasets or data subsets.

279 The models were developed and tested on the Exploratory dataset. Model
 280 hyperparameters were adjusted until the classification accuracies appeared to peak. The
 281 model architecture with the highest classification accuracy on the Exploratory dataset was
 282 trained, validated, and tested independently on the Confirmatory dataset. This means that

283 the model that was used to analyze the Confirmatory dataset was not trained on the
284 Exploratory dataset. The model architectures used for the timeline and plot image datasets
285 are shown in Figure 3.

286 Analysis

287 Results for the CNN architecture that resulted in the highest accuracy on the
288 Exploratory dataset are reported below. For every dataset tested, a one-sample two-tailed
289 *t*-test was used to compare the CNN accuracies against chance (33%). The Shapiro-Wilk test
290 was used to assess the normality for each dataset. When normality was assumed, the mean
291 accuracy for that dataset was compared against chance using Student's one-sample
292 two-tailed *t*-test. When normality could not be assumed, the median accuracy for that
293 dataset was compared against chance using Wilcoxon's Signed Rank test.

294 To determine the relative value of the three components of the eye movement data, the
295 data subsets were compared within the timeline and plot image data types. If classification
296 accuracies were lower when the data were batched into subsets, the component that was
297 removed was assumed to have some unique contribution that the model was using to inform
298 classification decisions. To determine the relative value of the contribution from each
299 component, the accuracies from each subset with one component of the data removed were
300 compared to the accuracies for the full dataset (XYP) using a one-way between-subjects
301 Analysis of Variance (ANOVA). To further evaluate the decodability of each component
302 independently, the accuracies from each subset containing only one component of the eye
303 movement data were compared within a separate one-way between-subjects ANOVA. All
304 post-hoc comparisons were corrected using Tukey's HSD.

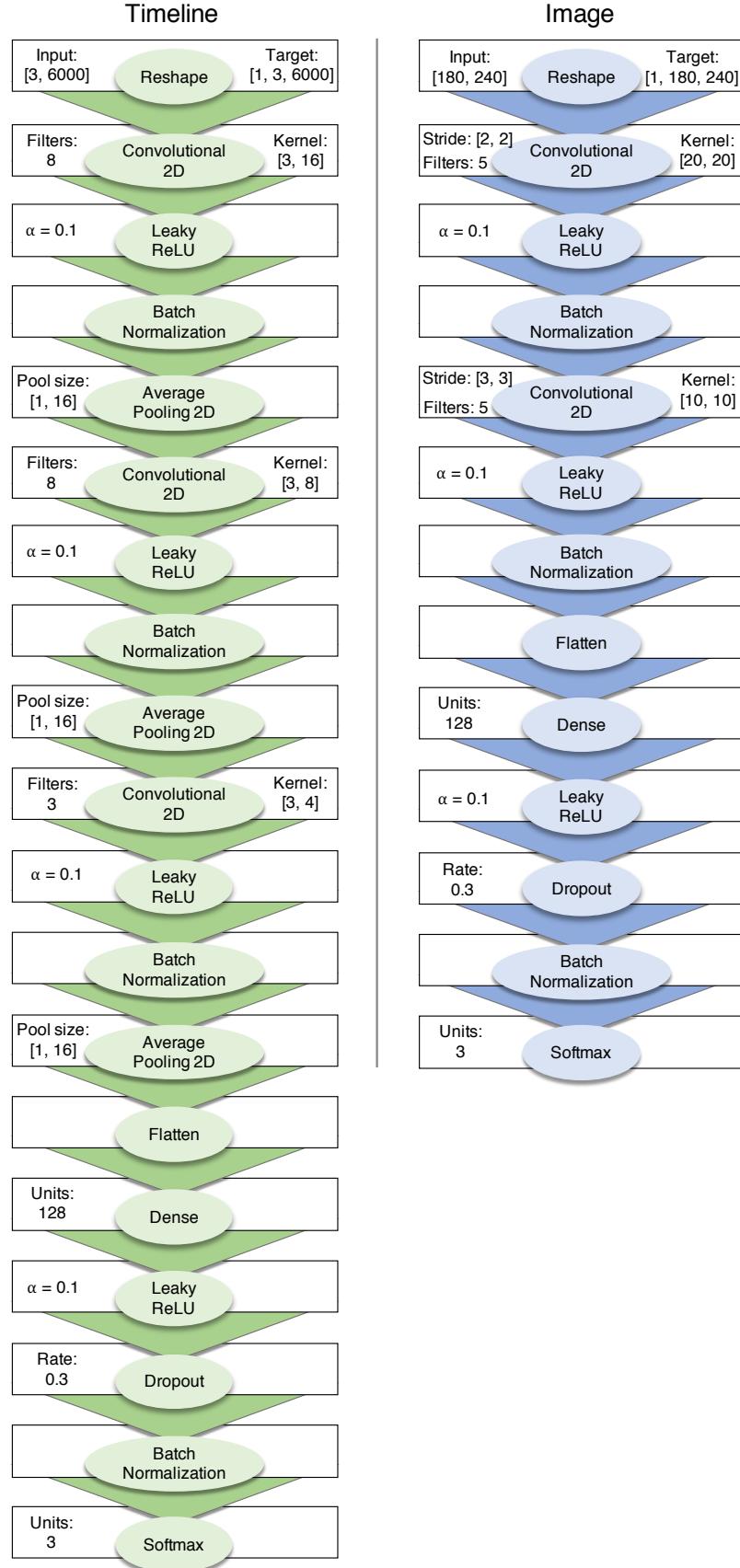


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm.

305

Results

306 Timeline Data Classification

307 **Exploratory.** Classification accuracies for the XYP timeline dataset were well above
 308 chance (chance = .33; $M = .526$, $SD = .018$; $t_9 = 34.565$, $p < .001$). Accuracies for
 309 classifications of the batched data subsets were all better than chance (see Figure 4). As
 310 shown in the confusion matrices displayed in Figure 5, the data subsets with lower overall
 311 classification accuracies almost always classified the Memorize condition at or below chance
 312 levels of accuracy. Misclassifications of the Memorize condition were split relatively evenly
 313 between the Search and Rate conditions.

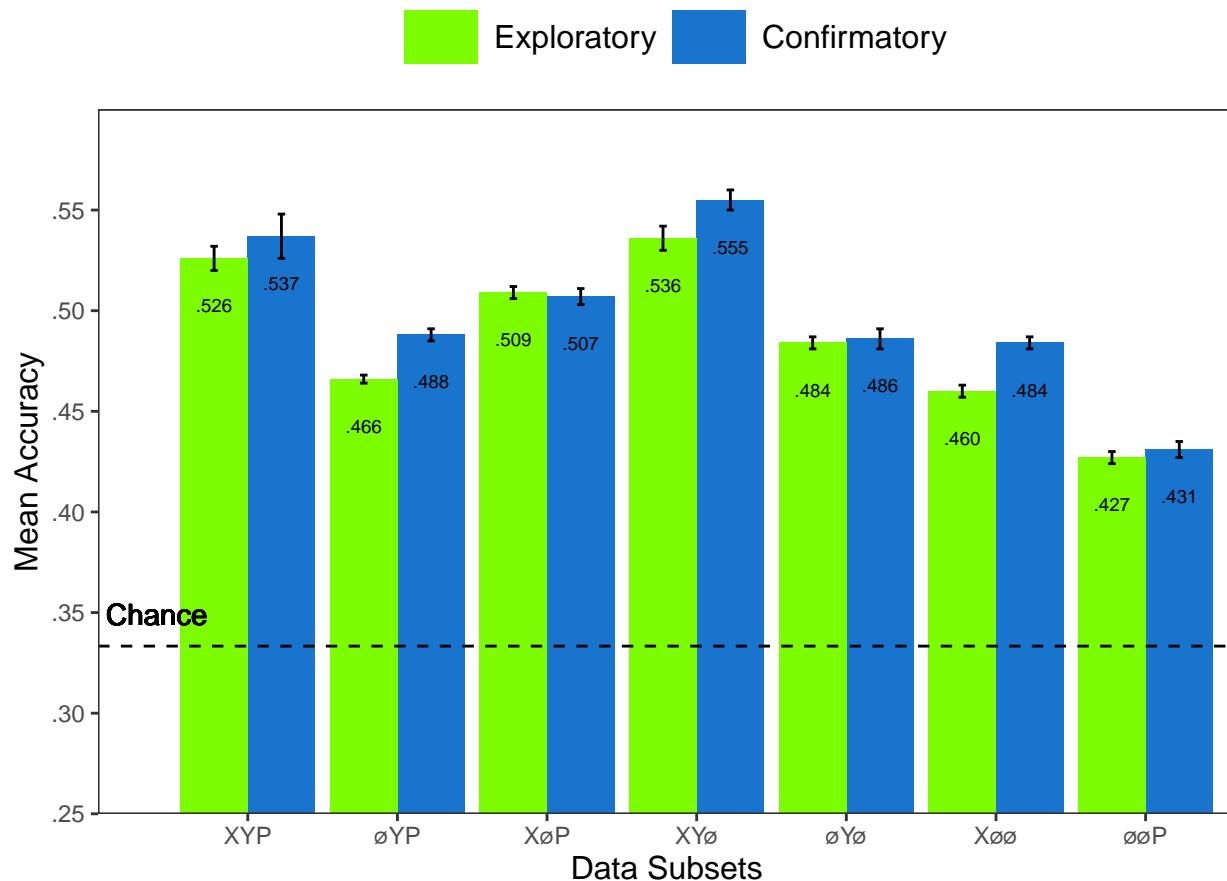


Figure 4. All of the data subsets were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

314

There was a difference in classification accuracy for the XYP dataset and the subsets

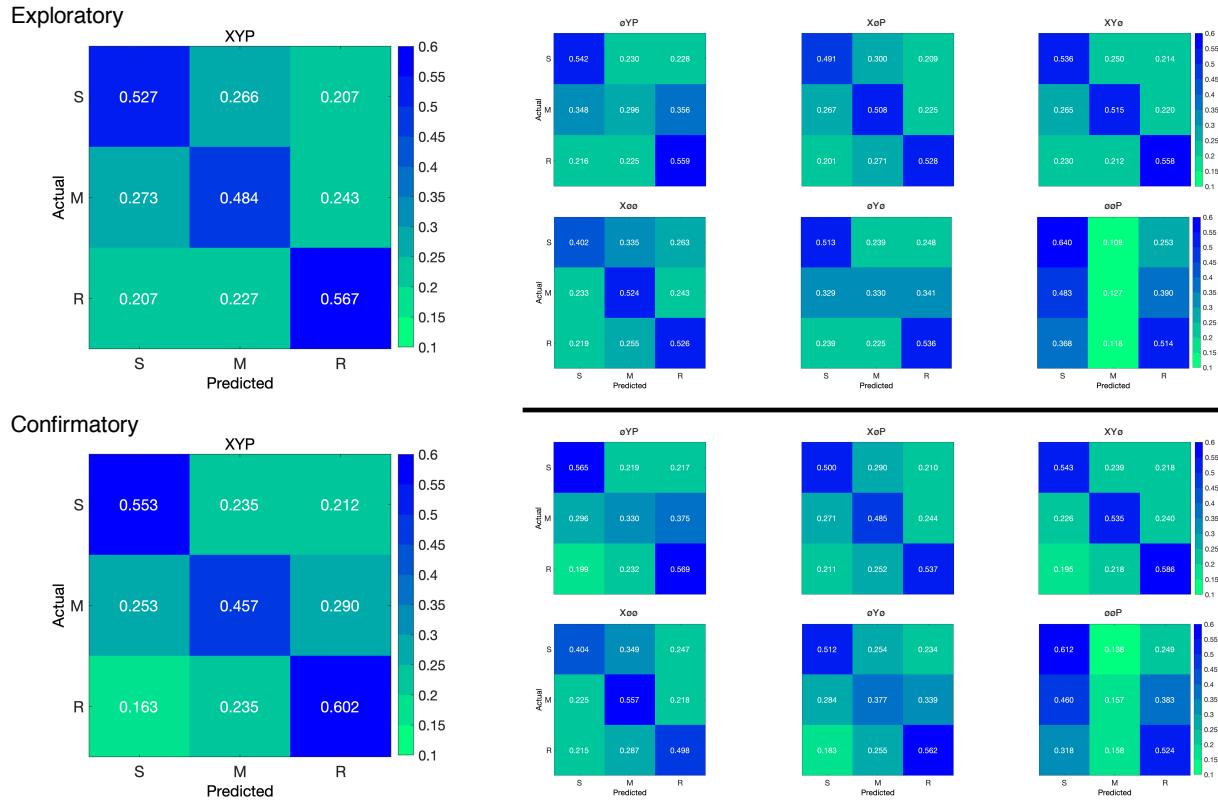


Figure 5. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

315 that had the pupil size, x-coordinate, and y-coordinate data systematically removed ($F_{3,36} =$
 316 $47.471, p < .001, \eta^2 = 0.798$). Post-hoc comparisons against the XYP dataset showed that
 317 classification accuracies were not affected by the removal of pupil size or y-coordinate data
 318 (see Table 2). The null effect present when pupil size was removed suggests that the pupil
 319 size data were not contributing unique information that was not otherwise provided by the x-
 320 and y-coordinates. A strict significance threshold of $\alpha = .05$ implies the same conclusion for
 321 the y-coordinate data, but the relatively low degrees of freedom ($df = 18$) and the borderline
 322 observed p -value ($p = .056$) afford the possibility that there exists a small effect. However,
 323 classification for the \emptyset YP subset was significantly lower than the XYP dataset, showing that
 324 the x-coordinate data were uniquely informative to the classification.

325 There was also a difference in classification accuracies for the X $\emptyset\emptyset$, \emptyset Y \emptyset , and $\emptyset\emptyset$ P
 326 subsets ($F_{2,27} = 75.145, p < .001, \eta^2 = 0.848$). Post-hoc comparisons showed that

Table 2
Timeline Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
XYP vs. \emptyset YP	9.420	< .001	5.210	< .001
XYP vs. X \emptyset P	2.645	.056	3.165	.016
XYP vs. XY \emptyset	1.635	.372	1.805	.288
X \emptyset \emptyset vs. \emptyset Y \emptyset	5.187	< .001	0.495	.874
X \emptyset \emptyset vs. \emptyset \emptyset P	12.213	< .001	10.178	< .001
\emptyset Y \emptyset vs. \emptyset \emptyset P	7.026	< .001	9.683	< .001

327 classification accuracy for the $\emptyset\emptyset$ P subset was lower than the X \emptyset \emptyset and \emptyset Y \emptyset subsets.
 328 Classification accuracy for the X \emptyset \emptyset subset was higher than the \emptyset Y \emptyset subset. Altogether,
 329 these findings suggest that pupil size data was the least uniquely informative to classification
 330 decisions, while the x-coordinate data was the most uniquely informative.

331 **Confirmatory.** Classification accuracies for the Confirmatory XYP timeline dataset
 332 were well above chance ($M = .537$, $SD = 0.036$, $t_9 = 17.849$, $p < .001$). Classification
 333 accuracies for the data subsets were also better than chance (see Figure 4). Overall, there
 334 was high similarity in the pattern of results for the Exploratory and Confirmatory datasets
 335 (see Figure 4). Furthermore, the general trend showing that pupil size was the least
 336 informative eye tracking data component was replicated in the Confirmatory dataset (see
 337 Table 2). Also in concordance with the Exploratory timeline dataset, the confusion matrices
 338 for these data revealed that the Memorize task was mis-classified more often than the Search
 339 and Rate tasks (see Figure 5).

340 To test the generalizability of the model architecture, classification accuracies for the
 341 XYP Exploratory and Confirmatory timeline datasets were compared. The Shapiro-Wilk
 342 test for normality indicated that the Exploratory ($W = 0.937$, $p = .524$) and Confirmatory
 343 ($W = 0.884$, $p = .145$) datasets were normally distributed, but Levene's test indicated that
 344 the variances were not equal, $F_{1,18} = 8.783$, $p = .008$. Welch's unequal variances t -test did
 345 not show a difference between the two datasets, $t_{13.045} = 0.907$, $p = .381$, Cohen's $d = 0.406$.

³⁴⁶ These findings indicate that the deep learning model decoded the Exploratory and
³⁴⁷ Confirmatory timeline datasets equally well, but the Confirmatory dataset classifications
³⁴⁸ were less consistent across training/test iterations (as indicated by the increase in standard
³⁴⁹ deviation).

³⁵⁰ **Plot Image Classification**

³⁵¹ **Exploratory.** Classification accuracies for the XYP plot image data were better
³⁵² than chance ($M = .436$, $SD = .020$, $p < .001$), but were less accurate than the classifications
³⁵³ for the XYP Exploratory timeline data ($t_{18} = 10.813$, $p < .001$). Accuracies for the
³⁵⁴ classifications for all subsets of the plot image data except the $\emptyset\emptyset P$ subset were better than
³⁵⁵ chance (see Figure 6). Following the pattern expressed by the timeline dataset, the confusion
³⁵⁶ matrices showed that the Memorize condition was misclassified more often than the other
³⁵⁷ conditions, and appeared to be equally mis-identified as a Search or Rate condition (see
³⁵⁸ Figure 7).

³⁵⁹ There was a difference in classification accuracy between the XYP dataset and the data
³⁶⁰ subsets ($F_{4,45} = 7.093$, $p < .001$, $\eta^2 = .387$). Post-hoc comparisons showed that compared to
³⁶¹ the XYP dataset, there was no effect of removing pupil size or the x-coordinates, but
³⁶² classification accuracy was worse when the y-coordinates were removed (see Table 3).

Table 3
Image Subset Comparisons

Comparison	Exploratory		Confirmatory	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
XYP vs. $\emptyset YP$	1.792	.391	1.623	.491
XYP vs. $X\emptyset P$	2.939	.039	4.375	< .001
XYP vs. $XY\emptyset$	0.474	.989	1.557	.532
$X\emptyset\emptyset$ vs. $\emptyset Y\emptyset$	0.423	.906	2.807	.204
$X\emptyset\emptyset$ vs. $\emptyset\emptyset P$	13.569	< .001	5.070	< .001
$\emptyset Y\emptyset$ vs. $\emptyset\emptyset P$	13.235	< .001	7.877	< .001

³⁶³ There was also a difference in classification accuracies between the $X\emptyset\emptyset$, $\emptyset Y\emptyset$, and
³⁶⁴ $\emptyset\emptyset P$ subsets (Levene's test: $F_{2,27} = 3.815$, $p = .035$; Welch correction for lack of

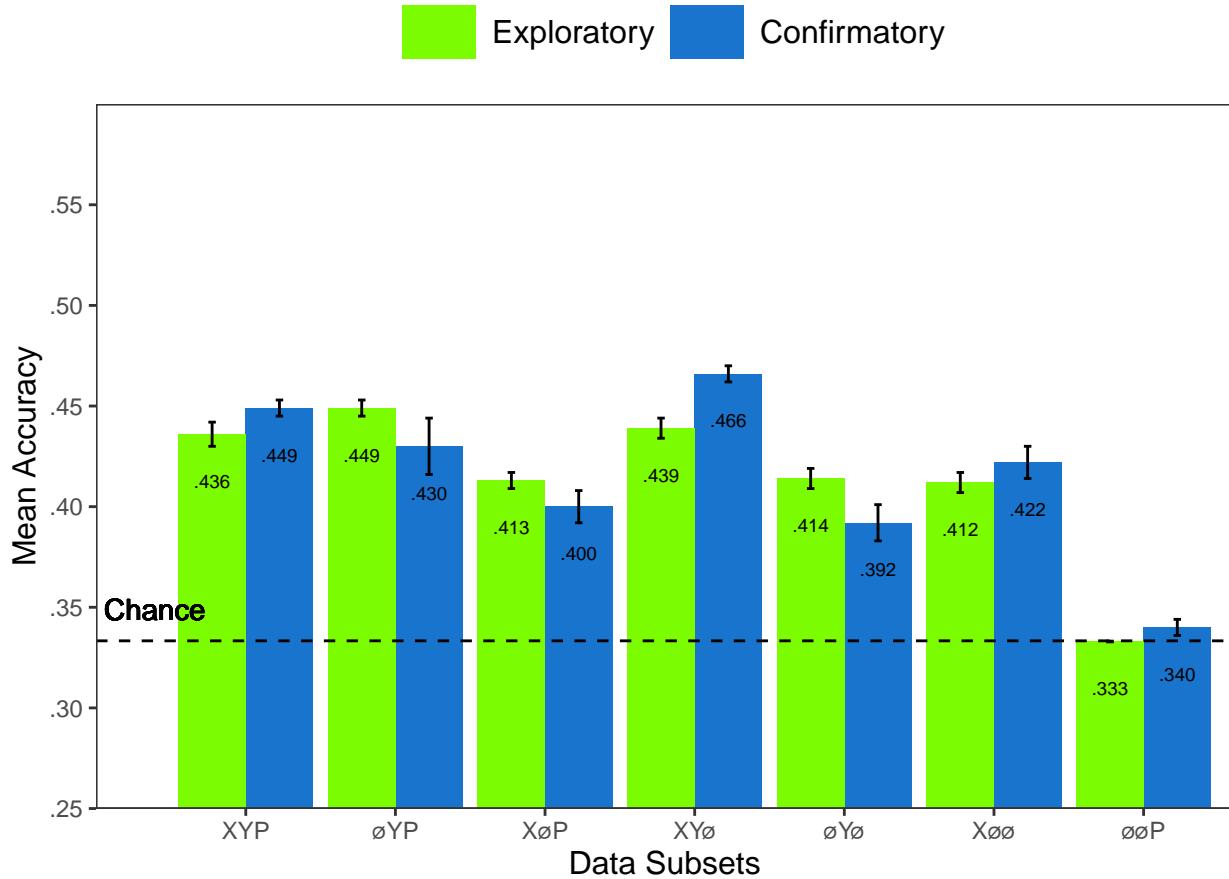


Figure 6. All of the data subsets except for the Exploratory ØØP dataset were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

homogeneity of variances: $F_{2,17.993} = 228.137, p < .001, \eta^2 = .899$). Post-hoc comparisons showed that there was no difference in classification accuracies for the XØØ and ØYØ subsets, but classification for the ØØP subset were less accurate than the XØØ and ØYØ subsets.

Confirmatory. Classification accuracies for the XYP confirmatory image dataset were well above chance ($M = .449, SD = 0.012, t_9 = 31.061, p < .001$), but were less accurate than the classifications of the confirmatory timeline dataset ($t_{18} = 11.167, p < .001$). Accuracies for classifications of the data subsets were also all better than chance (see Figure 6). The confusion matrices followed the pattern showing that the Memorize condition was mistaken most often, and was relatively equally mis-identified as a Search or Rate trial (see Figure 7). As with the timeline data, the general trend showing that pupil size data was

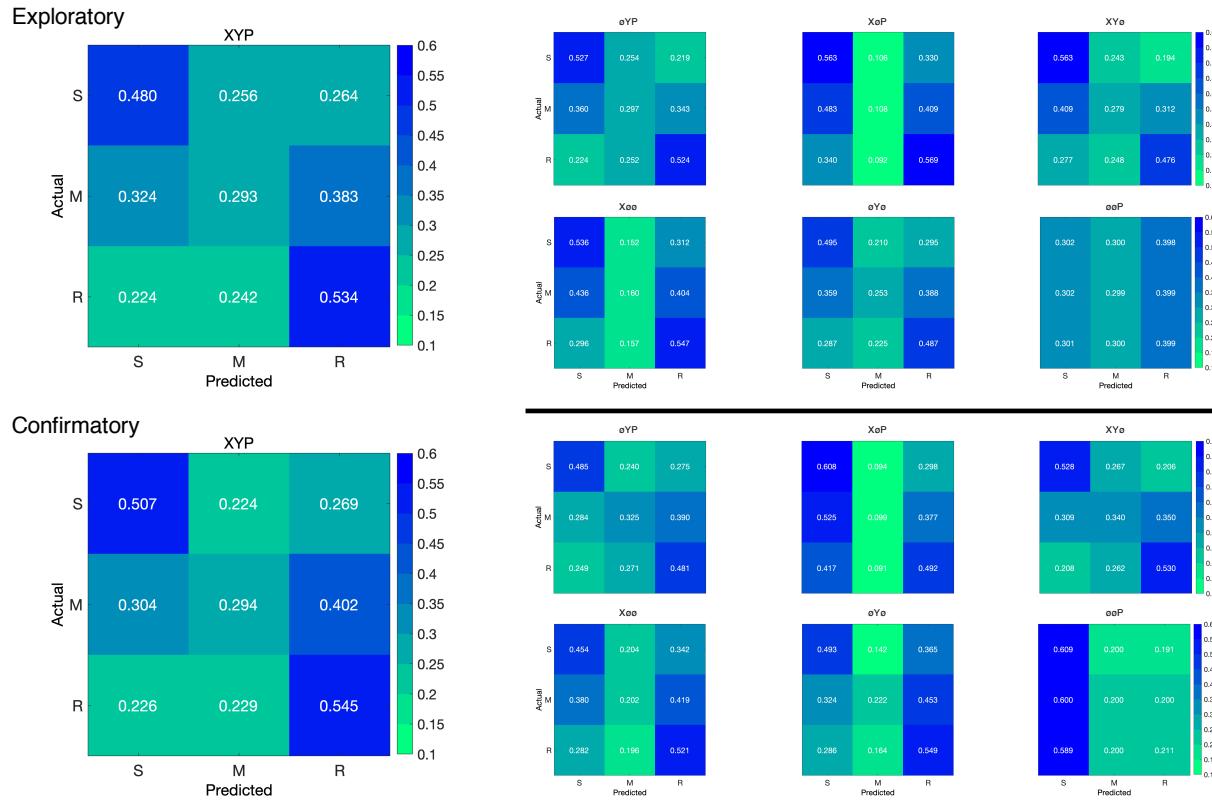


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

376 the least informative to the model was replicated in the Confirmatory dataset (see Table 3).

377 To test the generalizability of the model architecture, the classification accuracies for
 378 the XYP Exploratory and Confirmatory plot image datasets were compared. The
 379 independent samples *t*-test comparing the classification accuracies for the Exploratory and
 380 Confirmatory plot image datasets did not show a significant difference, $t_{18} = 1.777$, $p = .092$,
 381 Cohen's $d = 0.795$.

382 Discussion

383 The present study aimed to produce a practical and reliable example of a black box
 384 solution to the inverse Yarbus problem. To implement this solution, we classified raw
 385 timeline and minimally processed plot image data using a CNN model architecture. To our
 386 knowledge, this study was the first to provide a solution to determining mental state from

387 eye movement data using each of the following: (1) Non-aggregated eye tracking data (i.e.,
388 raw x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see
389 Figure 2), and (3) a black box CNN architecture. This study probed the relative predictive
390 value of the x-coordinate, y-coordinate, and pupil size components of the eye movement data
391 using a CNN. The CNN was able to decode the timeline and plot image data better than
392 chance, although only the timeline datasets were decoded with accuracies comparable to
393 other state-of-the-art approaches. Datasets with lower classification accuracies were not able
394 to differentiate the cognitive processes underlying the Memorize task from the cognitive
395 processes underlying the Search and Rate tasks. Decoding subsets of the data revealed that
396 pupil size was the least uniquely informative component of the eye movement data. This
397 pattern of findings was consistent between the Exploratory and Confirmatory datasets.

398 Although several aggregate eye movement features have been tested as task predictors,
399 to our knowledge, no other study has assessed the predictive value of the data format (viz.,
400 data in the format of a plot image). Our results suggest that although CNNs are robust
401 image classifiers, eye movement data is decoded in the standard timeline format more
402 effectively than in image format. This may be because the image data format contains less
403 decodable information than the timeline format. Over the span of the trial (six seconds), the
404 eye movements occasionally overlapped. When there was an overlap in the image data
405 format, the more recent data points overwrote the older data points. This resulted in some
406 information loss that did not occur when the data were represented in the raw timeline
407 format. Despite this loss of information, the plot image format was still decoded with better
408 than chance accuracy. To further examine the viability of classifying task from eye
409 movement image datasets, future research might consider representing the data in different
410 forms such as 3-dimensional data formats, or more complex color combinations capable of
411 representing overlapping data points.

412 When considering the superior performance of the timeline data (vs., plot image data),

413 we must also consider the differences in the model architectures. Because the structures of
414 the timeline and plot image data formats were different, the models decoding those data
415 structures also needed to be different. Both model architectures were optimized individually
416 on the Exploratory dataset before being tested on the Confirmatory dataset. For both
417 timeline and plot image formats, there was good replicability between the Exploratory and
418 Confirmatory datasets, demonstrating that these architectures performed similarly from
419 experiment to experiment. An appropriately tuned CNN should be capable of learning any
420 arbitrary function, but given that the upper bound for decodability of these datasets is
421 unknown, there is the possibility that a model architecture exists that is capable of
422 classifying the plot image data format more accurately than the model used to classify the
423 timeline data. Despite this possibility, the convergence of these findings with other studies
424 (see Table 1) suggests that the results of this study are approaching a ceiling for the
425 potential to solve the inverse Yarbus problem with eye movement data. Although the true
426 capacity to predict mental state from eye movement data is unknown, standardizing datasets
427 in the future could provide a point for comparison that can more effectively indicate which
428 methods are most effective at solving the inverse Yarbus problem.

429 In the current study, the Memorize condition was classified less accurately than the
430 Search and Rate conditions, especially for the datasets with lower overall accuracy. This
431 suggests that the eye movements associated with the Memorize task were potentially lacking
432 unique or informative features to decode. This means that eye movements associated with
433 the Memorize condition were interpreted as noise, or were sharing features of underlying
434 cognitive processes that were represented in the eye movements associated with the Search
435 and Rate tasks. Previous research (e.g., Król & Król, 2018) has attributed the inability to
436 differentiate one condition from the others to the overlapping of sub-features in the eye
437 movements between two tasks that are too subtle to be represented in the eye movement
438 data.

439 To more clearly understand how the different tasks influenced the decodability of the
440 eye movement data, additional analyses were conducted on the Exploratory and
441 Confirmatory timeline datasets (see Appendix). For the main supplementary analysis, the
442 data subsets were re-submitted to the CNN and re-classified as 2-category task sets. In
443 addition to the main supplementary analysis, the results from the primary analysis were
444 re-calculated from 3-category task sets to 2-category task sets. In the primary analyses, the
445 Memorize condition was predicted with the lowest accuracy, but mis-classifications of the
446 Search and Rate trials were most often categorized as Memorize. As a whole, this pattern of
447 results and the main supplementary analysis indicated a general bias for uncertain trials to
448 be categorized as Memorize. As expected, the main supplementary analysis also showed that
449 the 2-category task set that included only Search and Rate had higher accuracies than both
450 of the 2-category task sets that included the Memorize condition. The re-calculation analysis
451 generally replicated the pattern of results seen in the main supplementary analysis but with
452 larger variance, suggesting that including lower-accuracy trial types during model training
453 can decrease the consistency of classifier performance. Overall, the findings from this
454 supplemental analysis show that conclusions drawn from comparisons between approaches
455 that do not use the same task sets, or the same number of tasks, could be potentially
456 uninterpretable because the features underlying the task categories are interpreted differently
457 by the neural network algorithm.

458 When determining the relative contributions of the the eye movement features used in
459 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently the
460 least uniquely informative. When pupil size was removed from the Exploratory and
461 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,
462 XYP dataset). Furthermore, classification accuracy of the $\emptyset\emptyset\emptyset$ subset was the lowest of all
463 of the data subsets, and in one instance, was no better than chance. Although these findings
464 indicate that, in this case, pupil size was a relatively uninformative component of the eye
465 movement data, previous research has associated changes in pupil size as indicators of

466 working memory load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004),
467 arousal (Wang et al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The
468 results of the current study indicate that the changes in pupil size associated with these
469 underlying processes were not useful in delineating the tasks being classified (i.e., Search,
470 Memorize, Rate), potentially because these tasks did not evoke a reliable pattern of changes
471 in pupil size. Additionally, properties of the stimuli known to influence pupil size, such as
472 luminance and contrast, were not controlled in these datasets. Given that stimuli were
473 randomly assigned, there is the possibility that uncontrolled stimulus properties known to
474 affect pupil size impeded the CNN's capacity to detect patterns in the pupil size data.

475 The findings from the current study support the notion that black box CNNs are a
476 viable approach to determining task from eye movement data. In a recent review, Lukander
477 et al. (2017) expressed concern regarding the lack of generalizability of black box approaches
478 when decoding eye movement data. Overall, the current study showed a consistent pattern
479 of results for the XYP timeline and image datasets, but some minor inconsistencies in the
480 pattern of results for the x- and y- coordinate subset comparisons. These inconsistencies may
481 be a product of overlap in the cognitive processes underlying the three tasks. When the data
482 are batched into subsets, at least one dimension (i.e., x-coordinates, y-coordinates, or pupil
483 size) is removed, leading to a potential loss of information. When the data provide fewer
484 meaningful distinctions, finer-grained inferences are necessary for the tasks to be
485 distinguishable. As shown by Coco and Keller (2014), eye movement data can be more
486 effectively decoded when the cognitive processes underlying the tasks are explicitly
487 differentiable. While the cognitive processes distinguishing memorizing, searching, or rating
488 an image are intuitively different, the eye movements elicited from these cognitive processes
489 are not easily differentiated. To correct for potential mismatches between the distinctive
490 task-diagnostic features in the data and the level of distinctiveness required to classify the
491 tasks, future research could more definitively conceptualize the cognitive processes
492 underlying the task-at-hand.

493 Classifying mental state from eye movement data is often carried out in an effort to

494 advance technology to improve educational outcomes, strengthen the independence of

495 physically and mentally handicapped individuals, or improve HCI's (Koochaki &

496 Najafizadeh, 2018). Given the previous questions raised regarding the reliability and

497 generalizability of black-box CNN classification, the current study first tested models on an

498 exploratory dataset, then confirmed the outcome using a second independent dataset.

499 Overall, the findings of this study indicate that this black-box approach is capable of

500 producing a stable and generalizable outcome. Additionally, the supplementary analyses

501 showed that different task sets, or a different number of tasks, could lead the algorithm to

502 interpret features differently, which should be taken into account when comparing task

503 classification approaches. Future studies that incorporate features from the stimulus might

504 have the potential to surpass current state-of-the-art classification. According to Bulling,

505 Weichel, and Gellersen (2013), incorporating stimulus feature information into the dataset

506 may improve accuracy relative to decoding gaze location data and pupil size. Alternatively,

507 Borji and Itti (2014) suggested that accounting for salient features in the the stimulus might

508 leave little to no room for theoretically defined classifiers to consider mental state. Future

509 research should examine the potential for the inclusion of stimulus feature information in

510 addition to the eye movement data to boost black-box CNN classification accuracy of image

511 data beyond that of timeline data.

512

References

- 513 Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the
514 importance of spatial distribution, dynamics, and image features. *Neurocomputing*,
515 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- 516 Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task.
517 *Journal of Vision*, 14(3), 1–21. <https://doi.org/10.1167/14.3.29>
- 518 Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level
519 contextual cues from human visual behaviour. In *Proceedings of the SIGCHI
520 Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris,
521 France: ACM Press. <https://doi.org/10.1145/2470654.2470697>
- 522 Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye
523 movement control during active scene perception. *Journal of Vision*, 9(3), 1–15.
524 <https://doi.org/10.1167/9.3.6>
- 525 Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using
526 eye-movement features. *Journal of Vision*, 14(3), 1–18.
527 <https://doi.org/10.1167/14.3.11>
- 528 DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited.
529 *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- 530 Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict
531 observers' task from eye movement patterns. *Vision Research*, 62, 1–8.
532 <https://doi.org/10.1016/j.visres.2012.03.019>
- 533 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers'
534 task from eye movement patterns. *Vision Research*, 103, 127–142.

- 535 <https://doi.org/10.1016/j.visres.2014.08.014>
- 536 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).
537 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937.
538 <https://doi.org/10.1371/journal.pone.0064937>
- 539 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,
540 154(3756), 1583–1585. Retrieved from <http://www.jstor.org/stable/1720478>
- 541 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting
542 an observer's task using multi-fixation pattern analysis. In *Proceedings of the*
543 *Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290).
544 Safety Harbor, Florida: ACM Press. <https://doi.org/10.1145/2578153.2578208>
- 545 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the
546 dual-task paradigm as measured through behavioral and psychophysiological
547 responses. *Psychophysiology*, 41(2), 175–185.
548 <https://doi.org/10.1111/j.1469-8986.2004.00147.x>
- 549 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze Patterns.
550 In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).
551 <https://doi.org/10.1109/BIOCAS.2018.8584665>
- 552 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive processes
553 involved in the task from spatial eye-movement patterns. *Psychological Research*, 84,
554 245–258. <https://doi.org/10.1007/s00426-018-0996-5>
- 555 Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from Gaze
556 in Naturalistic Behavior: A Review. *International Journal of Mobile Human*
557 *Computer Interaction*, 9(4), 41–57. <https://doi.org/10.4018/IJMHCI.2017100104>

- 558 MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative
559 Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*,
560 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- 561 Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011).
562 Examining the influence of task set on eye movements and fixations. *Journal of*
563 *Vision*, 11(8), 1–15. <https://doi.org/10.1167/11.8.17>
- 564 Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and
565 counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*
566 (2006), 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- 567 Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., &
568 van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and
569 decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266.
570 <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- 571 Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus,
572 Eye Movements, and Vision. *I-Perception*, 1(1), 7–27. <https://doi.org/10.1068/i0382>
- 573 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018).
574 Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional
575 Face Task. *Frontiers in Neurology*, 9, 1029. <https://doi.org/10.3389/fneur.2018.01029>
- 576 Yarbus, A. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.
- 577 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep
578 Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K.
579 Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and*
580 *Visualizing Deep Learning* (pp. 243–252). Cham: Springer International Publishing.
581 https://doi.org/10.1007/978-3-030-28954-6_12

582

Appendix

583 Additional analyses were conducted in an attempt to clarify the effect of task on
 584 classification accuracy. These supplementary analyses were not seen as central to the current
 585 study, but could prove to be informative to researchers attempting to replicate or extend
 586 these findings in the future. The results from the primary analysis showed that classification
 587 accuracies were the lowest for the Memorize condition. To further understand why
 588 classification accuracy was lower for the Memorize condition than it was for the Search or
 589 Rate condition, the Exploratory and Confirmatory timeline datasets were systematically
 590 batched into subsets with the Search (S), Memorize (M), or Rate (R) condition removed (i.e.,
 591 \emptyset MR, S \emptyset R, SM \emptyset), and then run through the CNN classifier using the same methods as the
 592 primary analysis, but with only two classes.

593 All of the data subsets analyzed in this supplementary analysis were decoded with
 594 better than chance accuracy (see Figure 8a). The same pattern of results was observed in
 595 both the Exploratory and Confirmatory datasets. When the Memorize condition was
 596 removed, classification accuracy improved (see Table 4, Figure 8a). When the Rate condition
 597 was removed, classification was the worst. When the Memorize condition was included (i.e.,
 598 SM \emptyset and \emptyset MR), mis-classifications were biased toward Memorize, and the Memorize
 599 condition was more accurately predicted than the Search and Rate conditions (see Figure 9).

Table 4
Supplementary Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
\emptyset MR vs. S \emptyset R	3.248	.008	3.094	.012
\emptyset MR vs. SM \emptyset	2.875	.021	2.923	.018
S \emptyset R vs. SM \emptyset	6.123	< .001	6.017	< .001

600 The accuracies for all of the data subsets observed in the supplementary analysis were
 601 higher than the accuracies observed in the main analysis. Although there is a clear difference
 602 in accuracy, the primary analysis was classifying three categories (chance = .33) and the

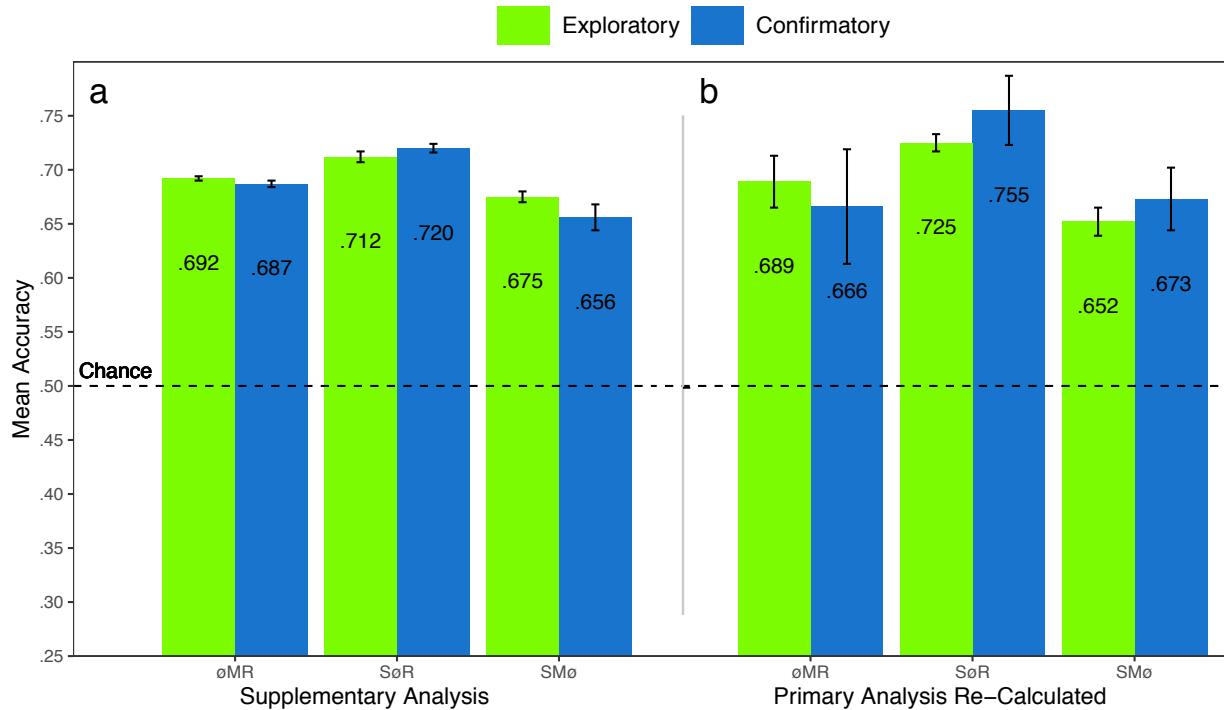


Figure 8. The graph represents the average accuracy reported for each subset of the Exploratory and Confirmatory timeline data for (a) the supplementary analysis, and the (b) re-calculated accuracies from the primary analysis. All of the data subsets were decoded at levels better than chance (.50). The error bars represent standard errors.

603 supplementary analysis was classifying two categories (chance = .50). Because the baseline
 604 chance performance was different for the primary and supplemental analyses, any conclusions
 605 drawn from a comparison of the results of analyses could be misleading. For this reason, we
 606 revisited the results from the primary analysis and re-calculated the predictions to be
 607 equivalent to a 50% chance threshold. Because the cross-validation scheme implemented by
 608 the DeLINEATE toolbox (<http://delineate.it>; Kuntzelman et al., under review) guaranteed
 609 an equal number of trials in the test set were assigned to each condition for each dataset, we
 610 were able to re-calculate 2-category predictions from the 3-category predictions presented in
 611 the confusion matrices from the primary analysis (see Figure 5). The predictions were
 612 re-calculated using the following formula: $\text{Prediction}_{(A,A,A \otimes C)} = \text{Prediction}_{(A,A,ABC)} /$
 613 $(\text{Prediction}_{(A,A,ABC)} + \text{Prediction}_{(A,C,ABC)})$. For example, accuracy for the Search
 614 classification for S \otimes R would be calculated with the following: $\text{Prediction}_{(S,S,S \otimes R)} =$
 615 $\text{Prediction}_{(S,S,SMR)} / (\text{Prediction}_{(S,S,SMR)} + \text{Prediction}_{(S,R,SMR)})$, where $\text{Prediction}_{(S,R,S \otimes R)}$ is

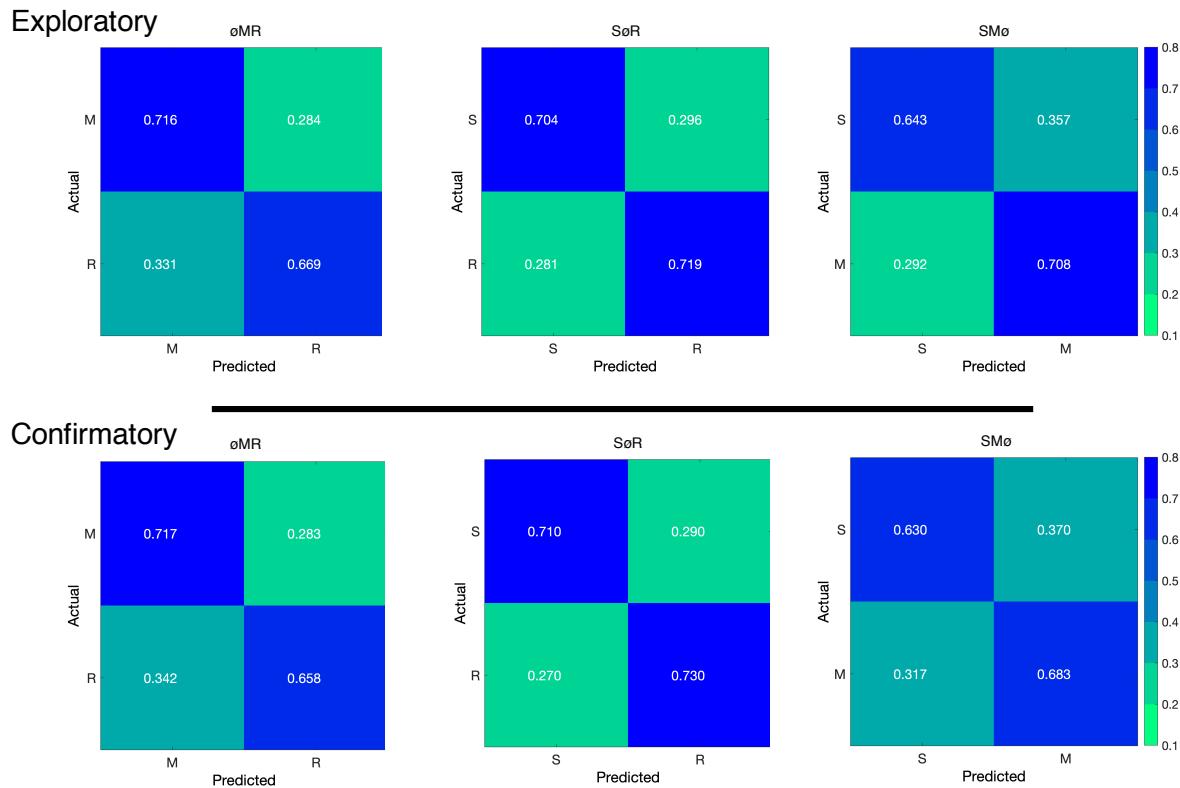


Figure 9. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

616 the ratio of Search trials that were misclassified as Rate.

617 The results for the re-calculated predictions followed a pattern similar to the main
 618 supplementary analysis (see Figure 8b). Looking back at the primary analysis, the
 619 3-category classifications predicted the Memorize conditions with the lowest accuracy (c.f.,
 620 Search and Rate conditions), and mis-classifications of the Search and Rate conditions were
 621 most often categorized as Memorize (see Figure 5). Because the Memorize condition was
 622 mis-classified more often than the other conditions in the primary analysis, the removal of
 623 the third class in the re-calculated SMø and øMR subsets resulted in a disproportionate
 624 amount of mis-classified Memorize trials being removed from those data subsets, somewhat
 625 eliminating the tendency to mis-classify Search and Rate trials as Memorize (see Figure 10).
 626 Nevertheless, the re-calculated SMø and øMR subsets were classified less accurately than
 627 SøR, just as in the main supplementary analysis.

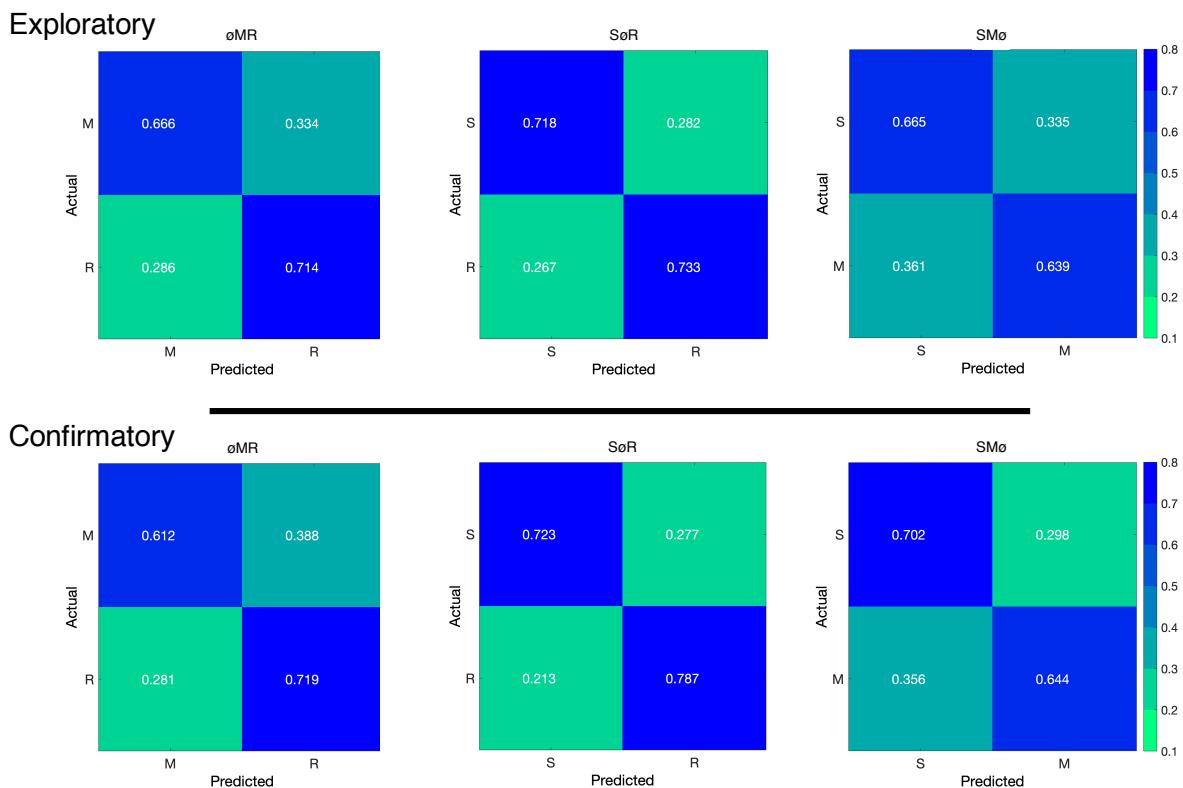


Figure 10. The confusion matrices represent a re-calculation of the classification accuracies for each category from the primary analysis. This re-calculation is meant to make the accuracies presented in the primary analysis (chance = .33) equivalent to the classification accuracies presented in the supplementary analysis (chance = .50).