

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 these model architectures has contributed to challenges in generalizing lab-trained CNNs to
24 applied contexts. In the current study, a CNN classifier was used to classify task from two
25 eye movement datasets (Exploratory and Confirmatory) in which participants searched,
26 memorized, or rated indoor and outdoor scene images. The Exploratory dataset was used to
27 tune the hyperparameters of the model, and the resulting model architecture was re-trained,
28 validated, and tested on the Confirmatory dataset. The data were formatted into timelines
29 (i.e., x-coordinate, y-coordinate, pupil size) and minimally processed images. To further
30 understand the informational value of each component of the eye movement data, the
31 timeline and image datasets were broken down into subsets with one or more components
32 systematically removed. Classification of the timeline data consistently outperformed the
33 image data. The Memorize condition was most often confused with Search and Rate. Pupil
34 size was the least uniquely informative component when compared with the x- and
35 y-coordinates. The general pattern of results for the Exploratory dataset was replicated in
36 the Confirmatory dataset. Overall, the present study provides a practical and reliable black
37 box solution to classifying task from eye movement data.

38 *Keywords:* deep learning, eye tracking, convolutional neural network, cognitive state,
39 endogenous attention

40

Introduction

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

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

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

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

92 Eye movements can only delineate tasks to the extent that the cognitive processes
93 underlying the tasks can be differentiated (Król & Król, 2018). Every task is associated with
94 a unique set of cognitive processes (Coco & Keller, 2014; Król & Król, 2018), but in some
95 cases, the cognitive processes for different tasks may produce indistinguishable eye movement
96 patterns. Others may define these terms differently, but for present purposes, our working
97 definitions are that cognitive "processes" are theoretical constructs that could be difficult to
98 isolate in practice, whereas a "task" is a more concrete/explicit set of goals and behaviors
99 imposed by the experimenter in an effort to operationalize one or more cognitive processes.
100 A "mental state," in contrast, is also a more theoretical term that is a bit more general and
101 could include goals and cognitive processes, but could also presumably encompass other
102 elements like mood or distraction.

103 To differentiate the cognitive processes underlying task-evoked eye movements, some
104 studies have chosen to classify tasks that rely on stimuli that prompt easily distinguishable
105 eye movements, such as reading text (e.g., Henderson, Shinkareva, Wang, Luke, & Olejarczyk,
106 2013). The eye movements elicited by salient stimulus features facilitate task classifications;
107 however, because these eye movements are the consequence of a feature (or features) inherent
108 to the stimulus rather than the task, it is unclear if these classifications are attributable to
109 the stimulus or a complex mental state (e.g., Henderson et al., 2013; Boisvert & Bruce,
110 2016). Additionally, the distinct nature of exogenously elicited eye movements prompts
111 decoding algorithms to prioritize these bottom-up patterns in the data over higher-level
112 top-down effects (Borji & Itti, 2014). This means that these models are identifying the type
113 of information that is being processed, but are not necessarily reflecting the mental state of
114 the individual observing the stimulus. Eye movements that are the product of bottom-up
115 processes have been reliably decoded, which is relevant for some HCI applications; however,
116 in our view such efforts do not fit the spirit of the inverse Yarbus problem, as most groups
117 seem to construe it. Namely, most attempts at addressing the inverse Yarbus problem are
118 concerned with decoding higher-level abstract mental operations that can be applied to

119 virtually any naturalistic image and are not directly dependent on specific structural
120 elements of the stimuli (e.g., the highly regular, linear patterns of written text).

121 Currently, there is not a clearly established upper limit to how well cognitive task can
122 be classified from eye movement data. Prior evidence has shown that the task-at-hand is
123 capable of producing distinguishable eye movement features such as the total scan path
124 length, total number of fixations, and the amount of time to the first saccade (Castelhano,
125 Mack, & Henderson, 2009; DeAngelus & Pelz, 2009). Decoding accuracies within the context
126 of determining task from eye movements typically range from chance performance to
127 relatively robust classification (see Table 1). In one case, Coco and Keller (2014) categorized
128 the same eye movement features used by Greene et al. (2012) with respect to the relative
129 contribution of latent visual or linguistic components of three tasks (visual search, name the
130 picture, name objects in the picture) with 84% accuracy (chance = 33%). While this
131 manipulation is reminiscent of other experiments relying on the bottom-up influence of
132 words and pictures (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016) the eye movements
133 in the Coco and Keller (2014) tasks can be attributed to the occurrence of top-down
134 attentional processes. A conceptually related follow-up to this study classified tasks along
135 two spatial and semantic dimensions, resulting in 51% classification accuracy (chance = 25%;
136 Król & Król, 2018). A closer look at these results showed that the categories within the
137 semantic dimension were consistently misclassified, suggesting that this level of distinction
138 may require a richer dataset, or a more powerful decoding algorithm. Altogether, there is no
139 measurable index of relative top-down or bottom-up influence, but this body of literature
140 suggests that the relative influence of top-down and bottom-up attentional processes may
141 have a role in determining the decodability of the eye movement data.

142 As shown in Table 1, when eye movement data are prepared for classification, fixation
143 and saccade statistics are typically aggregated along spatial or temporal dimensions,
144 resulting in variables such as fixation density or saccade amplitude (Castelhano et al., 2009;

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%)

¹⁴⁵ MacInnes et al., 2018; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). The¹⁴⁶ implementation of these statistical methods is meant to explicitly provide the decoding

147 algorithm with characteristics of the eye movement data that are representative of
148 theoretically relevant cognitive processes. For example, MacInnes et al. (2018) attempted to
149 provide an algorithm with data designed to be representative of inputs to the frontal eye
150 fields. In some instances, such as the case of Król and Król (2018), grounding the data using
151 theoretically driven aggregation methods may require sacrificing granularity in the dataset.
152 This means that aggregating the data has the potential to wash out certain fine-grained
153 distinctions that could otherwise be detected. Data structures of any kind can only be
154 decoded to the extent to which the data are capable of representing differences between
155 categories. Given that the cognitive processes underlying distinct tasks are often overlapping
156 (Coco & Keller, 2014), decreasing the granularity of the data may actually limit the potential
157 of the algorithm to make fine-grained distinctions between diagnostic components underlying
158 the tasks to be decoded.

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

The current study explored pragmatic solutions to the problem of classifying task from eye movement data by submitting **minimally processed** x-coordinate, y-coordinate, and pupil size data to a convolutional neural network (CNN) model. Instead of transforming the data into theoretically defined units, we allowed the network to learn meaningful patterns in the data on its own. CNNs have a natural propensity to develop low-level feature detectors similar to the primary visual cortex (e.g., Seeliger et al., 2018); for this reason, they are commonly implemented for image classification. In some cases, researchers have found success classifying data that natively exist in a timeline format by first transforming the data to an image-based format and then passing it to a deep neural network classifier (e.g., Bashivan, Rish, Yeasin, & Codella, 2016); however, it is not always obvious *a priori* which representation of a particular type of data is best-suited for neural network classifiers to be able to detect informative features, and the ideal representational format must be determined empirically. Thus, to test the possibility that image data might be better suited to the CNN classifier in our eye movement data as well, we also transformed our dataset from raw timelines into simple image representations and compared CNN-based classification of timeline data to that of image data. The image representations we generated also matched the eye movement trace images classically associated with the work of Yarbus (1967) and others, which were the original forays into this line of inquiry.

To our knowledge, no study has attempted to address the inverse Yarbus problem using any combination of the following methods: (1) Non-aggregated data, (2) image data format, and (3) a black-box CNN architecture. Given that CNN architectures are capable of learning features represented in raw data formats, and are well-suited to decoding multidimensional data that have a distinct spatial or temporal structure, we expected that a non-theoretically-constrained CNN architecture could be capable of decoding data at levels consistent with the current state of the art. Furthermore, despite evidence that black box approaches to the inverse Yarbus problem can impede generalizability (Lukander et al., 2017), we expected that when testing the approach on an entirely separate dataset, providing

200 the model with minimally processed data and the flexibility to identify the unique features
201 within each dataset would result in the replication of our initial findings.

202 **Method**

203 **Participants**

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

212 **Materials and Procedures**

213 Each participant viewed a series of indoor and outdoor scene images while carrying out
214 a search, memorization, or rating task. For the memorization task, participants were
215 instructed to memorize the image in anticipation of a forced choice recognition test. At the
216 end of each Memorize trial, the participants were prompted to indicate which of two images
217 was just presented. The two images were identical outside of a small change in the display
218 (e.g. object removed or added to the scene). For the rating task, participants were asked to
219 think about how they would rate the image on a scale from 1 (very unpleasant) to 7 (very
220 pleasant). The participants were prompted to provide a rating immediately after viewing the
221 image. For the search task, participants were instructed to find a small “Z” or “N” embedded
222 in the image. In reality, targets were not present in the images outside of a small subset of
223 images ($n = 5$) that were not analyzed but were included in the experiment design so
224 participants believed a target was always present. Trials containing the target were excluded

225 because search behavior was likely to stop if the target was found, adding considerable noise
226 to the eye movement data. For consistency between trial types, participants were prompted
227 to indicate if they found a “Z” or “N” at the end of each Search trial.

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

232 In both experiments, participants completed one mixed block of 120 trials (task cued
233 prior to each trial), or three uniform blocks of 40 trials (task cued prior to each block for a
234 total of 120 trials). Block type was assigned in counterbalanced order. When the blocks were
235 mixed, the trial types were randomly intermixed within the block. For uniform blocks, each
236 block consisted entirely of one of the three conditions (Search, Memorize, Rate), with block
237 types presented in random order. Each stimulus image was presented for 8 seconds. The
238 pictures were presented in color, with a size of 1024 x 768 pixels, subtending a visual angle of
239 $23.8^\circ \times 18.0^\circ$.

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

244 **Datasets**

245 On some trials, a probe was presented on the screen six seconds after the onset of the
246 trial, which required participants to fixate the probe once detected. To avoid confounds
247 resulting from the probe, only the first six seconds of the data for each trial was analyzed.
248 Trials that contained fewer than 6000 samples within the first six seconds of the trial were
249 excluded before analysis. For both datasets, the trials were pooled across participants. After

250 excluding trials, the Exploratory dataset consisted of 12,177 of the 16,740 total trials, and
 251 the Confirmatory dataset consisted of 9,301 of the 10,395 total trials.

252 The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling
 253 time point in the trial were used as inputs to the deep learning classifier. These data were
 254 also used to develop plot image datasets that were classified separately from the raw timeline
 255 datasets. For the plot image datasets, the timeline data for each trial were converted into
 256 scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data
 257 point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location
 258 of the dot, pupil size was used to determine the relative size of the dot, and shading of the
 259 dot was used to indicate the time-course of the eye movements throughout the trial. The
 260 background of the plot images and first data point were white. Each subsequent data point
 261 was one shade darker than the previous data point until the final data point was reached.
 262 The final data point was black. For standardization, pupil size was divided by 10, and one
 263 unit was added. The plots were sized to match the dimensions of the data collection monitor
 264 (1024 x 768 pixels) and then shrunk to (240 x 180 pixels) in an effort to reduce the
 265 dimensionality of the data.

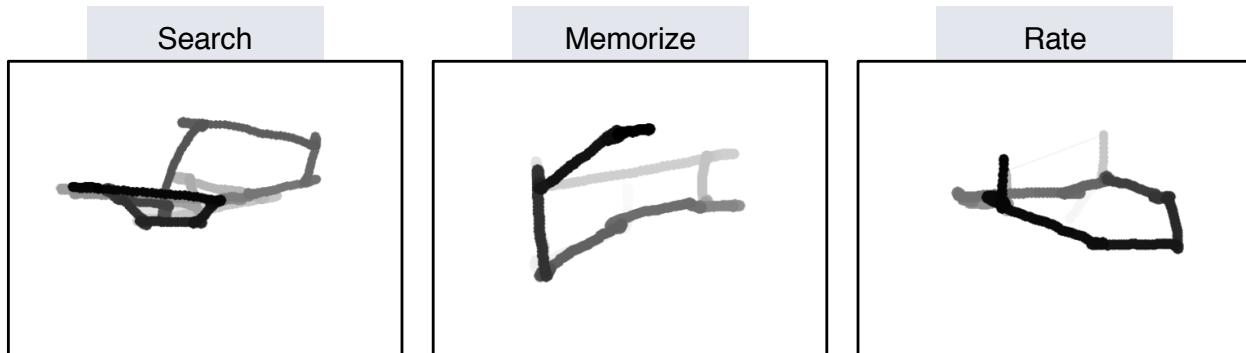


Figure 1. Each trial was represented as an image. Each sample collected within the trial was plotted as a dot in the image. Pupil size was represented by the size of the dot. The time course of the eye movements was represented by the gradual darkening of the dot over time.

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

²⁶⁹ x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image
²⁷⁰ datasets were batched into subsets that excluded one of the components (i.e., XYØ, XØP,
²⁷¹ ØYP), or contained only one of the components (i.e., XØØ, ØYØ, ØØP). For the timeline
²⁷² datasets, this means that the columns to be excluded in each data subset were replaced with
²⁷³ zeros. The data were replaced with zeros because removing the columns would change the
²⁷⁴ structure of the data. The same systematic batching process was carried out for the image
²⁷⁵ dataset. See Figure 2 for an example of each of these image data subsets.

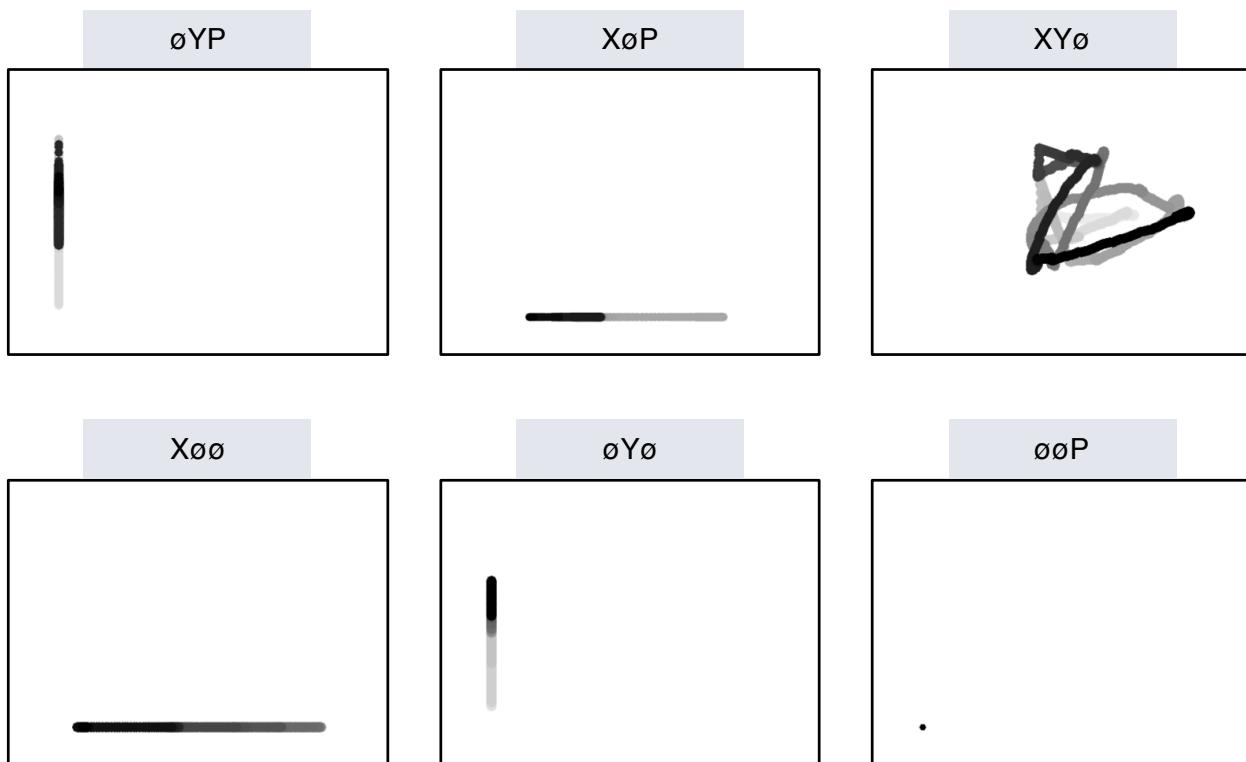


Figure 2. Plot images were used to represent data subsets that excluded one component of the eye movement data (i.e., XYØ, XØP, ØYP) or contained only one component (i.e., XØØ, ØYØ, ØØP). As with the trials in the full XYP dataset, the time course of the eye movements was represented by the shading of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

²⁷⁶ Classification

²⁷⁷ Deep CNN model architectures were implemented to classify the trials into Search,
²⁷⁸ Memorize, or Rate categories. Because CNNs act as a digital filter sensitive to the number of
²⁷⁹ features in the data, the differences in the structure of the timeline and image data formats
²⁸⁰ necessitated separate CNN model architectures. The model architectures were developed

281 with the intent of establishing a generalizable approach to classifying task from eye
282 movement data.

283 The development of these models was not guided by any formal theoretical
284 assumptions regarding the patterns or features likely to be extracted by the classifier. Like
285 many HCI models, the development of these models followed general intuitions concerned
286 with building a model architecture capable of transforming the data inputs into an
287 interpretable feature set that would not overfit the dataset. The models were developed
288 using version 0.3b of the DeLINEATE toolbox, which operates over a Keras backend
289 (<http://delineate.it>; Kuntzman et al., *in press*). Each training/test iteration randomly split
290 the data so that 70% of the trials were allocated to training, 15% to validation, and 15% to
291 testing. (This approach achieves essentially the same benefit of a more traditional k-fold
292 cross-validation approach insofar as it allows all data to be used as both training and test
293 without double-dipping; however, by resampling the data instead of using strict fold
294 divisions, we can sidestep the issue of how to incorporate a validation set into the k-fold
295 approach.) Training of the model was stopped when validation accuracy did not improve
296 over the span of 100 epochs. Once the early stopping threshold was reached, the resulting
297 model was tested on the held-out test data. This process was repeated 10 times for each
298 model, resulting in 10 classification accuracy scores for each model. The resulting accuracy
299 scores were used for the comparisons against chance and other datasets or data subsets.

300 The models were developed and tested on the Exploratory dataset. Model
301 hyperparameters were adjusted until the classification accuracies on the test data appeared
302 to peak, with no obvious evidence of excessive overfitting during the training process. The
303 model architecture with the highest classification accuracy on the Exploratory dataset was
304 trained, validated, and tested independently on the Confirmatory dataset. This means that
305 the model that was used to analyze the Confirmatory dataset was not trained on the
306 Exploratory dataset. For all of the analyses that excluded one or more components of the

307 eye movement data (e.g., XYØ, XØP, ØYP, and so on), new models were trained for each
308 data subset (i.e., data subset analyses did not use the model that had already been trained
309 on the full XYP dataset). The model architectures used for the timeline and plot image
310 datasets are shown in Figure 3, with some additional details on the architecture
311 hyperparameters in the figure caption.

312 **Analysis**

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

320 To determine the independent contributions of the three components of the eye
321 movement data, the data subsets were compared within the timeline and plot image data
322 types. If classification accuracies were lower when the data were batched into subsets, the
323 component that was removed was assumed to have some unique contribution that the model
324 was using to inform classification decisions. To determine the uniqueness of the contribution
325 from each component, the accuracies from each subset with one component of the data
326 removed were compared to the accuracies for the full dataset (XYP) using a one-way
327 between-subjects Analysis of Variance (ANOVA). To further evaluate the decodability of
328 each component independently, the accuracies from each subset containing only one
329 component of the eye movement data were compared within a separate one-way
330 between-subjects ANOVA. All post-hoc comparisons were corrected using Tukey's HSD.

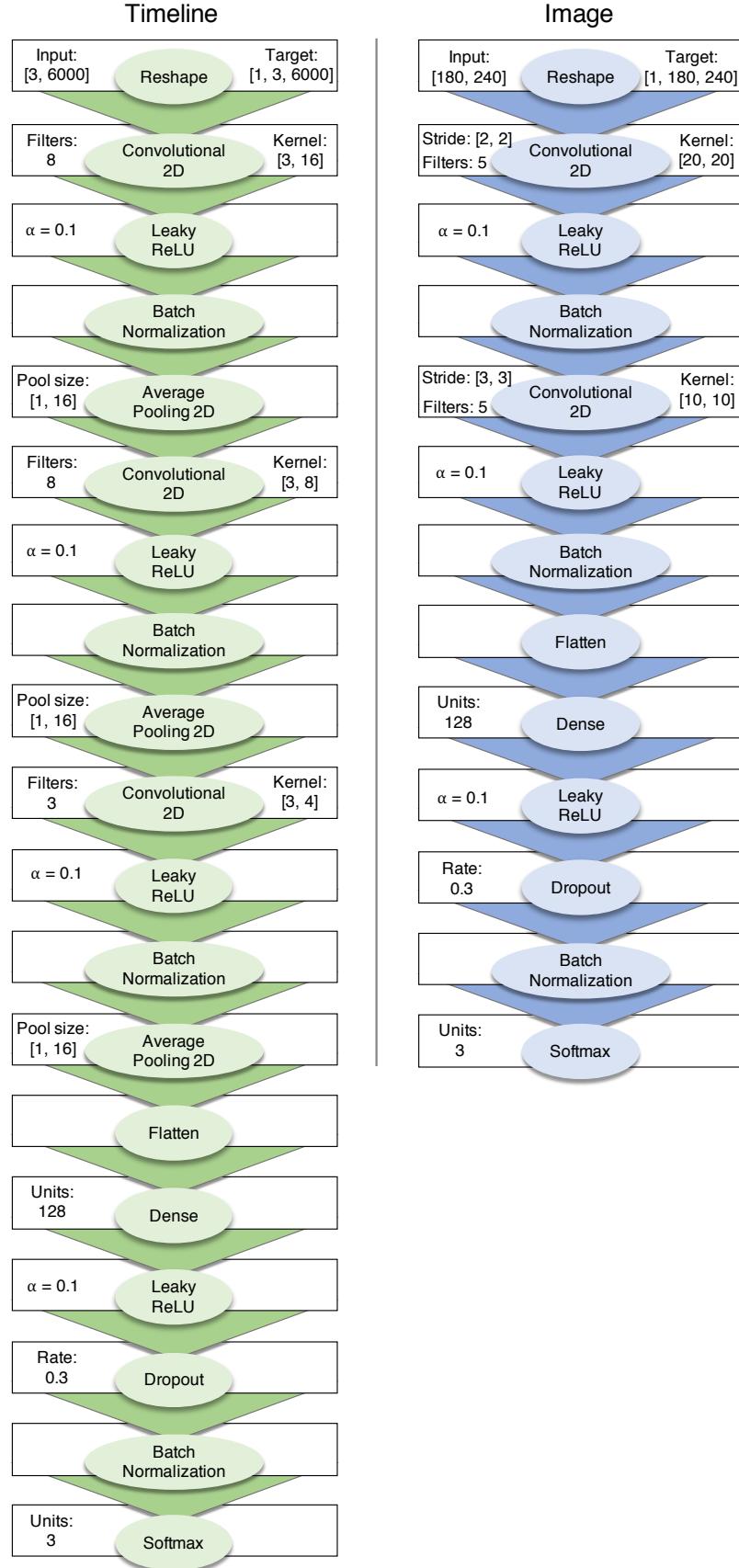


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm. Optimizer parameters were initial learning rate = 0.005, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 0.1$. The timeline model had 16,946 trainable parameters (29,998 total); the image model had 18,525 trainable parameters (18,827 total).

331

Results

332 Timeline Data Classification

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

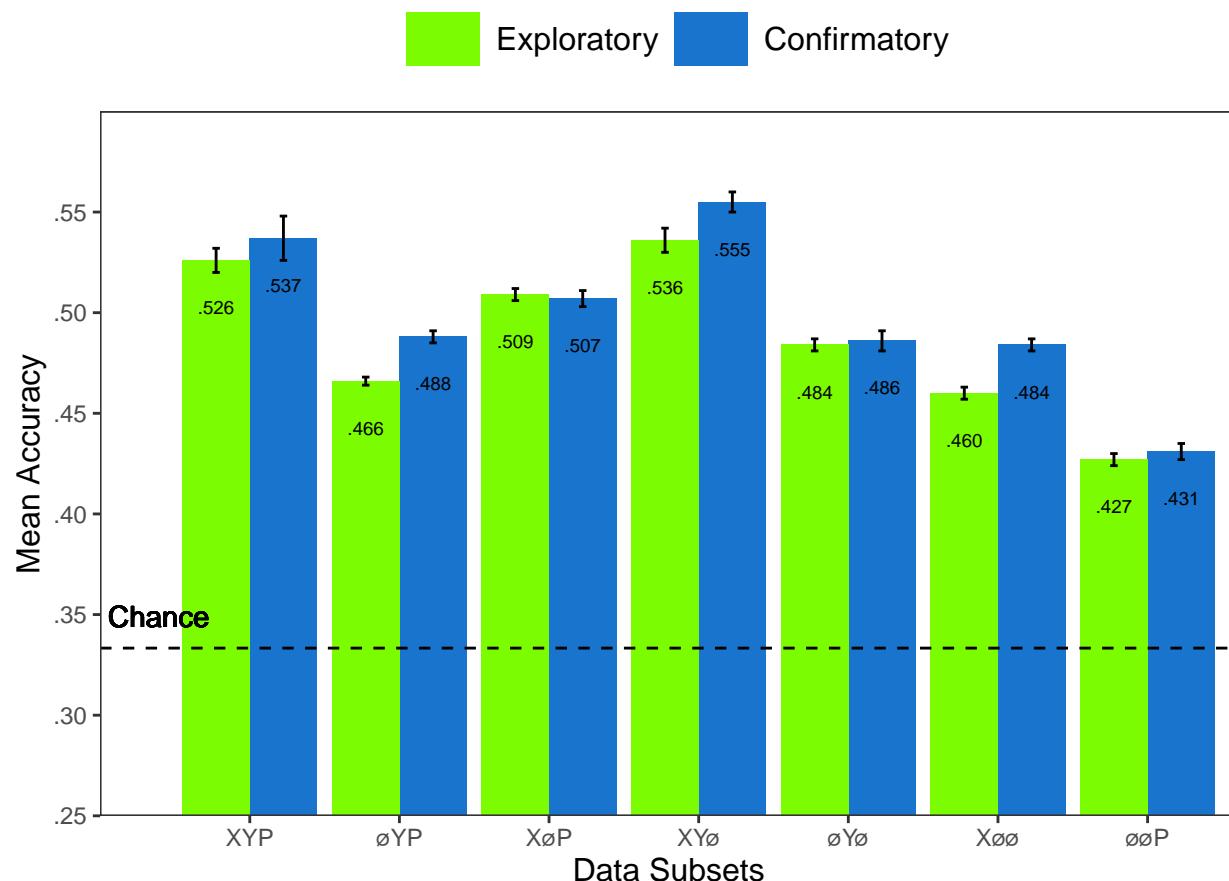


Figure 4. All of the data subsets were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

340

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

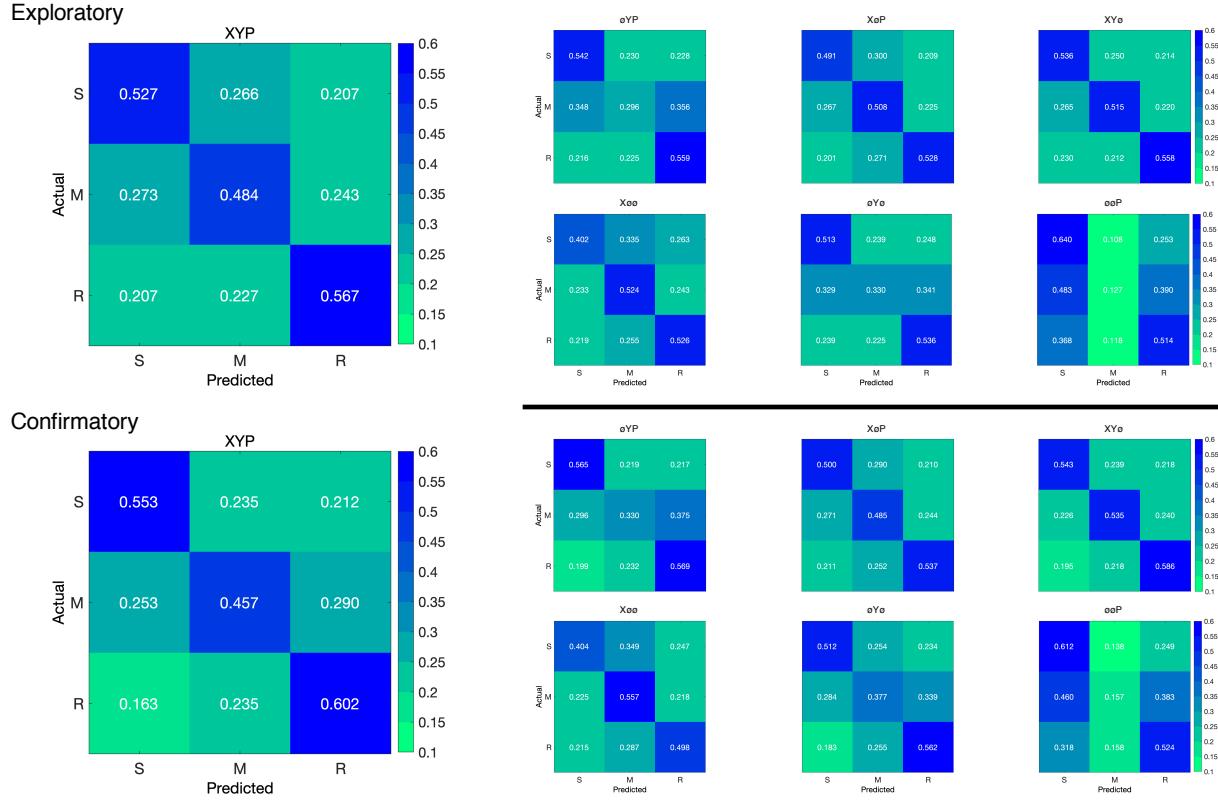


Figure 5. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

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

There was also a difference in classification accuracies for the X $\emptyset\emptyset$, \emptyset Y \emptyset , and $\emptyset\emptyset$ P 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

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

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

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

³⁷² These findings indicate that the deep learning model decoded the Exploratory and
³⁷³ Confirmatory timeline datasets equally well, but the Confirmatory dataset classifications
³⁷⁴ were less consistent across training/test iterations (as indicated by the increase in standard
³⁷⁵ deviation).

³⁷⁶ Plot Image Classification

³⁷⁷ **Exploratory.** Classification accuracies for the XYP plot image data were better
³⁷⁸ than chance ($M = .436$, $SD = .020$, $p < .001$), but were less accurate than the classifications
³⁷⁹ for the XYP Exploratory timeline data ($t_{18} = 10.813$, $p < .001$). Accuracies for the
³⁸⁰ classifications for all subsets of the plot image data except the $\emptyset\emptyset P$ subset were better than
³⁸¹ chance (see Figure 6). Following the pattern expressed by the timeline dataset, the confusion
³⁸² matrices showed that the Memorize condition was misclassified more often than the other
³⁸³ conditions, and appeared to be equally mis-identified as a Search or Rate condition (see
³⁸⁴ Figure 7).

³⁸⁵ There was a difference in classification accuracy between the XYP dataset and the data
³⁸⁶ subsets ($F_{4,45} = 7.093$, $p < .001$, $\eta^2 = .387$). Post-hoc comparisons showed that compared to
³⁸⁷ the XYP dataset, there was no effect of removing pupil size or the x-coordinates, but
³⁸⁸ 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

³⁸⁹ There was also a difference in classification accuracies between the $X\emptyset\emptyset$, $\emptyset Y\emptyset$, and
³⁹⁰ $\emptyset\emptyset P$ subsets (Levene's test: $F_{2,27} = 3.815$, $p = .035$; Welch correction for lack of

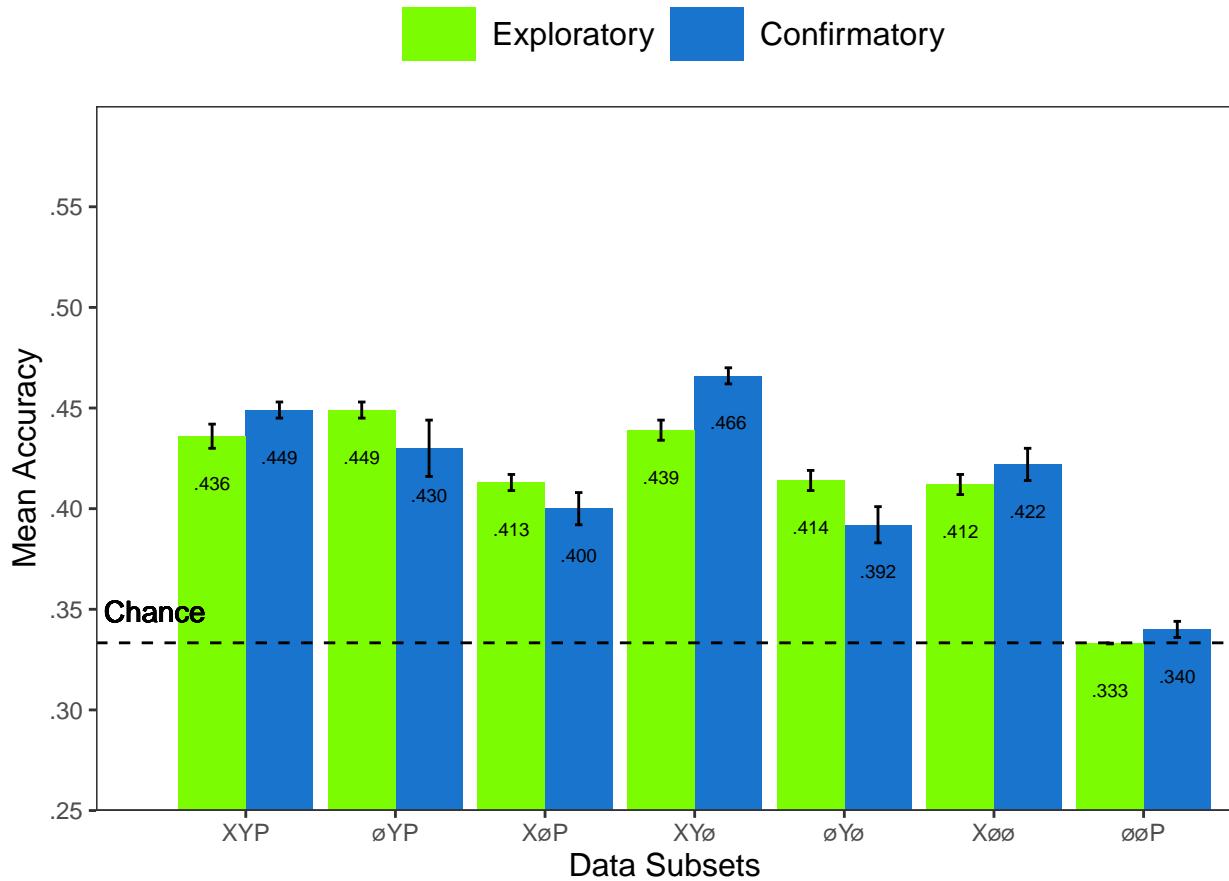


Figure 6. All of the data subsets except for the Exploratory ØØP dataset were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

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

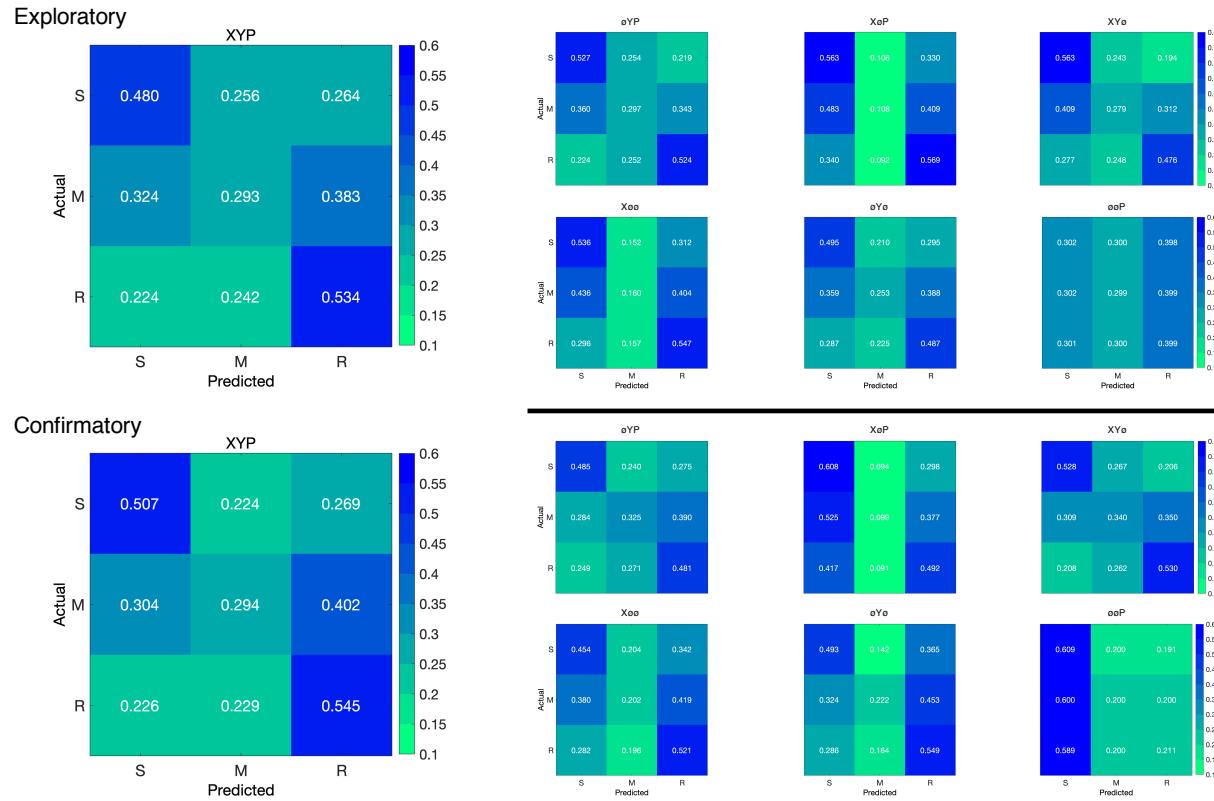


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

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

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

408 Discussion

409 The present study aimed to produce a practical and reliable example of a black box
 410 solution to the inverse Yarbus problem. To implement this solution, we classified raw
 411 timeline and minimally processed plot image data using a CNN model architecture. To our
 412 knowledge, this study was the first to provide a solution to determining task from eye

413 movement data using each of the following: (1) Non-aggregated eye tracking data (i.e., raw
414 x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see Figure 2),
415 and (3) a black box CNN architecture. This study probed the **independent contributions** of
416 the x-coordinate, y-coordinate, and pupil size components of the eye movement data using a
417 CNN. The CNN was able to decode the timeline and plot image data better than chance,
418 although only the timeline datasets were decoded with accuracies comparable to other
419 state-of-the-art approaches. Datasets with lower classification accuracies were not able to
420 differentiate the cognitive processes underlying the Memorize task from the cognitive
421 processes underlying the Search and Rate tasks. Decoding subsets of the data revealed that
422 pupil size was the least uniquely informative component of the eye movement data. This
423 pattern of findings was consistent between the Exploratory and Confirmatory datasets.

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

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

439 we must also consider the differences in the model architectures. Because the structures of
440 the timeline and plot image data formats were different, the models decoding those data
441 structures also needed to be different. Both model architectures were optimized individually
442 on the Exploratory dataset before being tested on the Confirmatory dataset. For both
443 timeline and plot image formats, there was good replicability between the Exploratory and
444 Confirmatory datasets, demonstrating that these architectures performed similarly from
445 experiment to experiment. An appropriately tuned CNN should be capable of learning any
446 arbitrary function, but given that the upper bound for decodability of these datasets is
447 unknown, there is the possibility that a model architecture exists that is capable of
448 classifying the plot image data format more accurately than the model used to classify the
449 timeline data. Despite this possibility, the convergence of these findings with other studies
450 (see Table 1) suggests that the results of this study are approaching a ceiling for the
451 potential to solve the inverse Yarbus problem with eye movement data. We attempted to
452 replicate some of those other studies' methods on our own dataset, but were only able to do
453 so with the methods of Coco and Keller (2014), due to lack of publicly available code or
454 incompatibility with our data; for Coco and Keller's methods, we did not achieve
455 better-than-chance classification in our data. Although the true capacity to predict mental
456 state from eye movement data is unknown, standardizing datasets in the future could
457 provide a point for comparison that can more effectively indicate which methods are most
458 effective at solving the inverse Yarbus problem. As a gesture towards this goal, we have
459 made the data and code from the present study publicly available at: <https://osf.io/dyq3t>.

460 In the current study, the Memorize condition was classified less accurately than the
461 Search and Rate conditions, especially for the datasets with lower overall accuracy. This
462 suggests that the eye movements associated with the Memorize task were potentially lacking
463 unique or informative features to decode. This means that eye movements associated with
464 the Memorize condition were interpreted as noise, or were sharing features of underlying
465 cognitive processes that were represented in the eye movements associated with the Search

466 and Rate tasks. Previous research (e.g., Król & Król, 2018) has attributed the inability to
467 differentiate one condition from the others to the overlapping of sub-features in the eye
468 movements between two tasks that are too subtle to be represented in the eye movement
469 data.

470 To more clearly understand how the different tasks influenced the decodability of the
471 eye movement data, additional analyses were conducted on the Exploratory and
472 Confirmatory timeline datasets (see Appendix). For the main supplementary analysis, the
473 data subsets were re-submitted to the CNN and re-classified as 2-category task sets. In
474 addition to the main supplementary analysis, the results from the primary analysis were
475 re-calculated from 3-category task sets to 2-category task sets. In the primary analyses, the
476 Memorize condition was predicted with the lowest accuracy, but mis-classifications of the
477 Search and Rate trials were most often categorized as Memorize. As a whole, this pattern of
478 results and the main supplementary analysis indicated a general bias for uncertain trials to
479 be categorized as Memorize. As expected, the main supplementary analysis also showed that
480 the 2-category task set that included only Search and Rate had higher accuracies than both
481 of the 2-category task sets that included the Memorize condition. The re-calculation analysis
482 generally replicated the pattern of results seen in the main supplementary analysis but with
483 larger variance, suggesting that including lower-accuracy trial types during model training
484 can decrease the consistency of classifier performance. Overall, the findings from this
485 supplemental analysis show that conclusions drawn from comparisons between approaches
486 that do not use the same task sets, or the same number of tasks, could be potentially
487 uninterpretable because the features underlying the task categories are interpreted differently
488 by the neural network algorithm.

489 When determining the **unique** contributions of the the eye movement features used in
490 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently the
491 least uniquely informative. When pupil size was removed from the Exploratory and

492 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,
493 XYP dataset). Furthermore, classification accuracy of the ØØP subset was the lowest of all
494 of the data subsets, and in one instance, was no better than chance. Although these findings
495 indicate that, in this case, pupil size was a relatively uninformative component of the eye
496 movement data, previous research has associated changes in pupil size as indicators of
497 working memory load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004),
498 arousal (Wang et al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The
499 results of the current study indicate that the changes in pupil size associated with these
500 underlying processes were not useful in delineating the tasks being classified (i.e., Search,
501 Memorize, Rate), potentially because these tasks did not evoke a reliable pattern of changes
502 in pupil size. Additionally, properties of the stimuli known to influence pupil size, such as
503 luminance and contrast, were not controlled in these datasets. Given that stimuli were
504 randomly assigned, there is the possibility that uncontrolled stimulus properties known to
505 affect pupil size impeded the CNN's capacity to detect patterns in the pupil size data.

506 The findings from the current study support the notion that black box CNNs are a
507 viable approach to determining task from eye movement data. In a recent review, Lukander
508 et al. (2017) expressed concern regarding the lack of generalizability of black box approaches
509 when decoding eye movement data. Overall, the current study showed a consistent pattern
510 of results for the XYP timeline and image datasets, but some minor inconsistencies in the
511 pattern of results for the x- and y- coordinate subset comparisons. These inconsistencies may
512 be a product of overlap in the cognitive processes underlying the three tasks. When the data
513 are batched into subsets, at least one dimension (i.e., x-coordinates, y-coordinates, or pupil
514 size) is removed, leading to a potential loss of information. When the data provide fewer
515 meaningful distinctions, finer-grained inferences are necessary for the tasks to be
516 distinguishable. As shown by Coco and Keller (2014), eye movement data can be more
517 effectively decoded when the cognitive processes underlying the tasks are explicitly
518 differentiable. While the cognitive processes distinguishing memorizing, searching, or rating

519 an image are intuitively different, the eye movements elicited from these cognitive processes
520 are not easily differentiated. To correct for potential mismatches between the distinctive
521 task-diagnostic features in the data and the level of distinctiveness required to classify the
522 tasks, future research could more definitively conceptualize the cognitive processes
523 underlying the task-at-hand.

524 Classifying mental state from eye movement data is often carried out in an effort to
525 advance technology to improve educational outcomes, strengthen the independence of
526 physically and mentally handicapped individuals, or improve HCI's (Koochaki &
527 Najafizadeh, 2018). Given the previous questions raised regarding the reliability and
528 generalizability of black-box CNN classification, the current study first tested models on an
529 exploratory dataset, then confirmed the outcome using a second independent dataset.
530 Overall, the findings of this study indicate that this black-box approach is capable of
531 producing a stable and generalizable outcome. Additionally, the supplementary analyses
532 showed that different task sets, or a different number of tasks, could lead the algorithm to
533 interpret features differently, which should be taken into account when comparing task
534 classification approaches. Future studies that incorporate features from the stimulus might
535 have the potential to surpass current state-of-the-art classification. According to Bulling,
536 Weichel, and Gellersen (2013), incorporating stimulus feature information into the dataset
537 may improve accuracy relative to decoding gaze location data and pupil size. Alternatively,
538 Borji and Itti (2014) suggested that accounting for salient features in the the stimulus might
539 leave little to no room for theoretically defined classifiers to consider mental state. Future
540 research should examine the potential for the inclusion of stimulus feature information in
541 addition to the eye movement data to boost black-box CNN classification accuracy of image
542 data beyond that of timeline data.

543

References

- 544 Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2016). Learning Representations from
545 EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv:1511.06448 [Cs]*.
546 Retrieved from <http://arxiv.org/abs/1511.06448>
- 547 Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the
548 importance of spatial distribution, dynamics, and image features. *Neurocomputing*,
549 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- 550 Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task.
551 *Journal of Vision*, 14(3), 1–21. <https://doi.org/10.1167/14.3.29>
- 552 Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level
553 contextual cues from human visual behaviour. In *Proceedings of the SIGCHI
554 Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris,
555 France: ACM Press. <https://doi.org/10.1145/2470654.2470697>
- 556 Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye
557 movement control during active scene perception. *Journal of Vision*, 9(3), 1–15.
558 <https://doi.org/10.1167/9.3.6>
- 559 Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using
560 eye-movement features. *Journal of Vision*, 14(3), 1–18.
561 <https://doi.org/10.1167/14.3.11>
- 562 DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited.
563 *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- 564 Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict
565 observers' task from eye movement patterns. *Vision Research*, 62, 1–8.

566 <https://doi.org/10.1016/j.visres.2012.03.019>

567 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers'
568 task from eye movement patterns. *Vision Research*, 103, 127–142.

569 <https://doi.org/10.1016/j.visres.2014.08.014>

570 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).
571 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937.

572 <https://doi.org/10.1371/journal.pone.0064937>

573 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,
574 154(3756), 1583–1585. Retrieved from <http://www.jstor.org/stable/1720478>

575 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting
576 an observer's task using multi-fixation pattern analysis. In *Proceedings of the*
577 *Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290).
578 Safety Harbor, Florida: ACM Press. <https://doi.org/10.1145/2578153.2578208>

579 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the
580 dual-task paradigm as measured through behavioral and psychophysiological
581 responses. *Psychophysiology*, 41(2), 175–185.

582 <https://doi.org/10.1111/j.1469-8986.2004.00147.x>

583 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze Patterns.
584 In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).
585 <https://doi.org/10.1109/BIOCAS.2018.8584665>

586 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive processes
587 involved in the task from spatial eye-movement patterns. *Psychological Research*, 84,
588 245–258. <https://doi.org/10.1007/s00426-018-0996-5>

- 589 Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from Gaze
590 in Naturalistic Behavior: A Review. *International Journal of Mobile Human*
591 *Computer Interaction*, 9(4), 41–57. <https://doi.org/10.4018/IJMHCI.2017100104>
- 592 MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative
593 Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*,
594 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- 595 Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011).
596 Examining the influence of task set on eye movements and fixations. *Journal of*
597 *Vision*, 11(8), 1–15. <https://doi.org/10.1167/11.8.17>
- 598 Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and
599 counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*
600 (2006), 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- 601 Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., &
602 van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and
603 decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266.
604 <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- 605 Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus,
606 Eye Movements, and Vision. *I-Perception*, 1(1), 7–27. <https://doi.org/10.1068/i0382>
- 607 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018).
608 Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional
609 Face Task. *Frontiers in Neurology*, 9, 1029. <https://doi.org/10.3389/fneur.2018.01029>
- 610 Yarbus, A. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.
- 611 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep

612 Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K.
613 Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and*
614 *Visualizing Deep Learning* (pp. 243–252). Cham: Springer International Publishing.
615 https://doi.org/10.1007/978-3-030-28954-6_12

616

Appendix

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

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

Table 4
Supplementary Subset Comparisons

Comparison	Exploratory		Confirmatory	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
\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

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

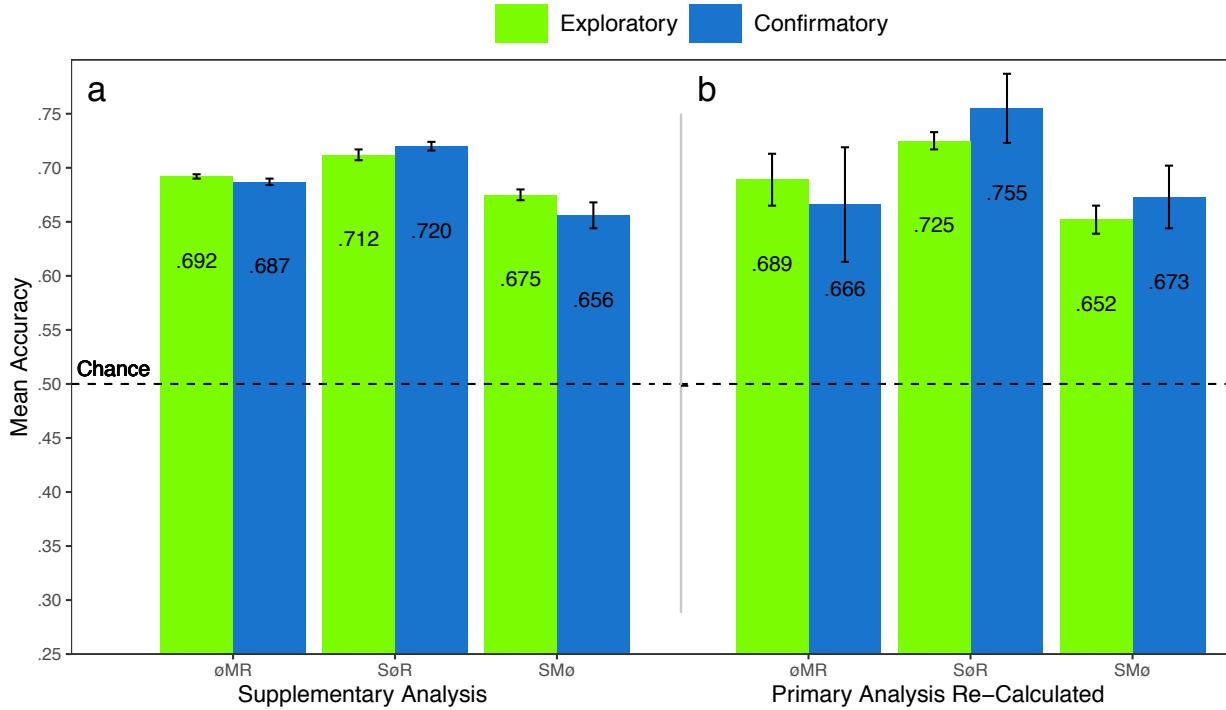


Figure 8. The graph represents the average accuracy reported for each subset of the Exploratory and Confirmatory timeline data for (a) the supplementary analysis, and the (b) re-calculated accuracies from the primary analysis. All of the data subsets were decoded at levels better than chance (.50). The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

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

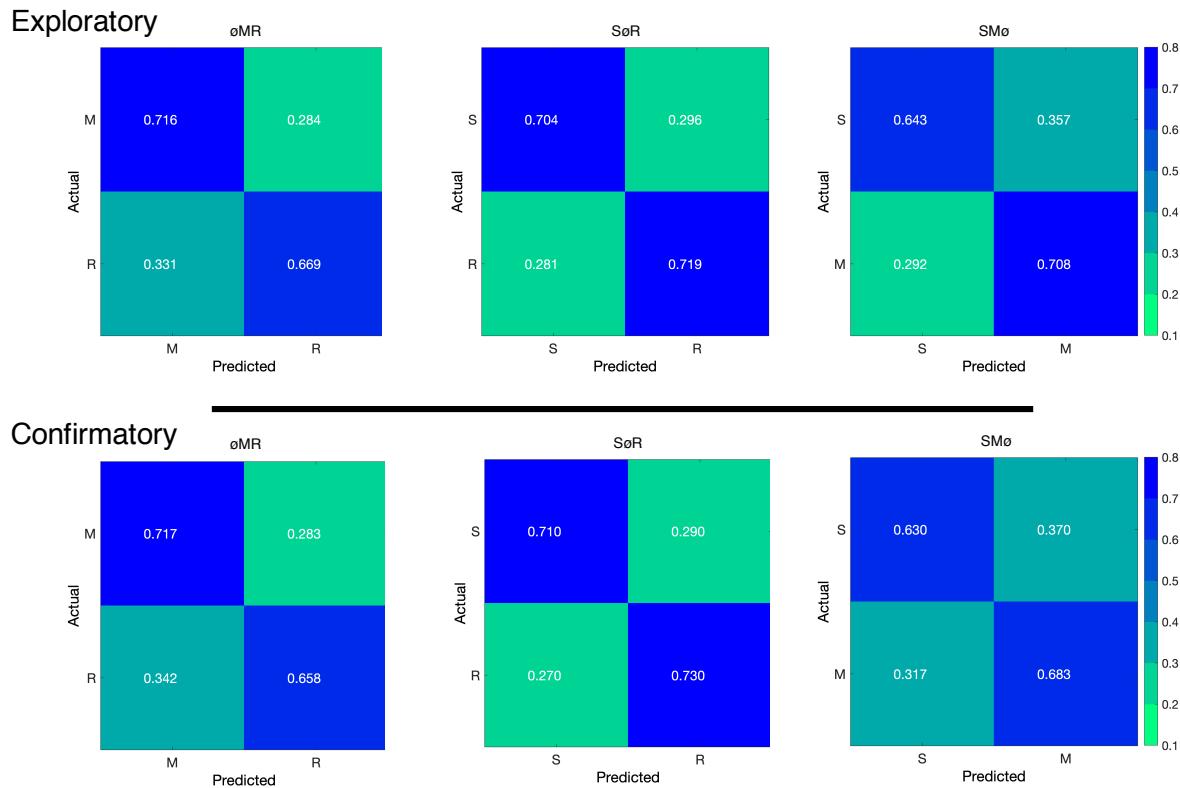


Figure 9. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

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

651 The results for the re-calculated predictions followed a pattern similar to the main
 652 supplementary analysis (see Figure 8b). Looking back at the primary analysis, the
 653 3-category classifications predicted the Memorize conditions with the lowest accuracy (c.f.,
 654 Search and Rate conditions), and mis-classifications of the Search and Rate conditions were
 655 most often categorized as Memorize (see Figure 5). Because the Memorize condition was
 656 mis-classified more often than the other conditions in the primary analysis, the removal of
 657 the third class in the re-calculated SMø and øMR subsets resulted in a disproportionate
 658 amount of mis-classified Memorize trials being removed from those data subsets, somewhat
 659 eliminating the tendency to mis-classify Search and Rate trials as Memorize (see Figure 10).
 660 Nevertheless, the re-calculated SMø and øMR subsets were classified less accurately than
 661 SøR, just as in the main supplementary analysis.

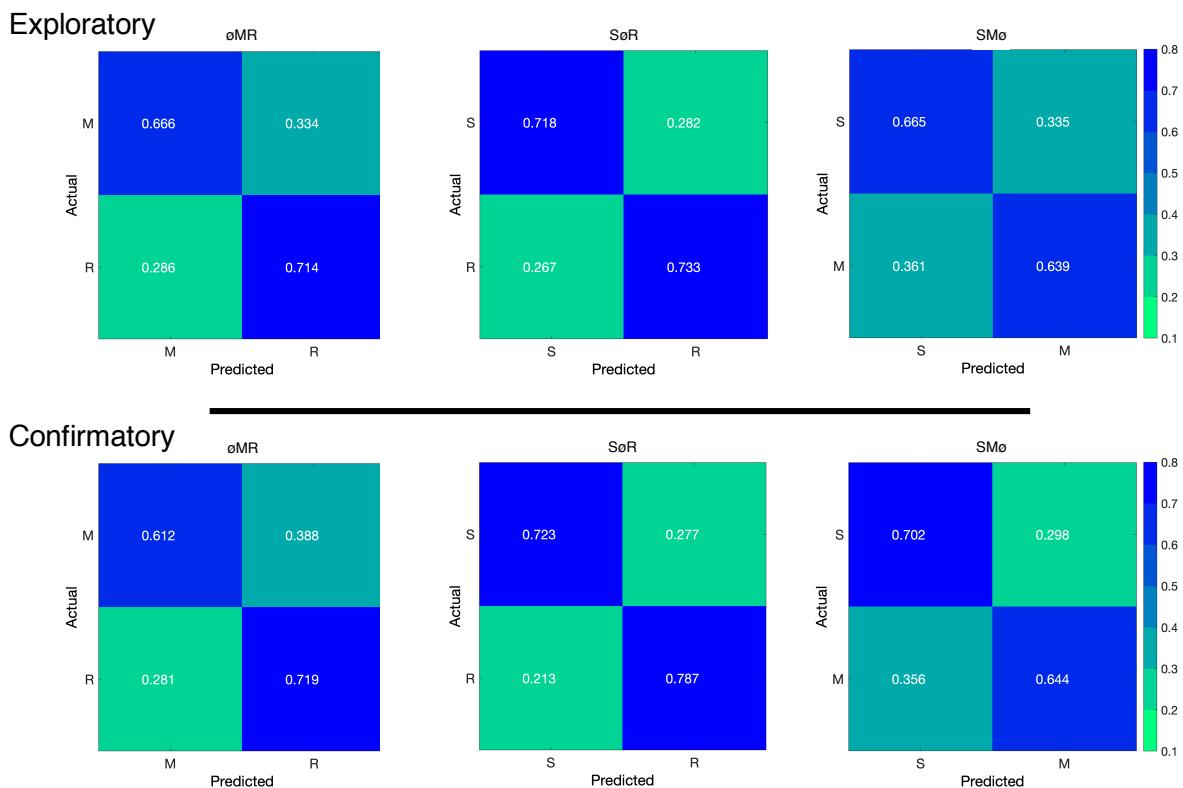


Figure 10. The confusion matrices represent a re-calculation of the classification accuracies for each category from the primary analysis. This re-calculation is meant to make the accuracies presented in the primary analysis (chance = .33) equivalent to the classification accuracies presented in the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