

Compare Results

Old File:

z_deepeye_draft1.pdf

34 pages (7.67 MB)

8/13/20, 8:25:34 PM

versus

New File:

z_deepeye.pdf

34 pages (7.27 MB)

8/20/20, 7:16:53 PM

Total Changes

721

Content

| | |
|------------|--------------|
| 452 | Replacements |
| 118 | Insertions |
| 111 | Deletions |

Styling and Annotations

| | |
|-----------|-------------|
| 29 | Styling |
| 11 | Annotations |

[Go to First Change \(page 2\)](#)

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. Additionally, work done to develop the analysis approach was supported by NSF/EPSCoR grant #1632849 and NIH grant GM130461 awarded to MRJ and colleagues.

¹⁰ Correspondence concerning this article should be addressed to Zachary J. Cole, 238
¹¹ Burnett Hall, Lincoln, NE 68588-0308. E-mail: z@neurophysicole.com

12

Abstract

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

35

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

36

endogenous attention

37

Word count: 7260

38

Introduction

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

49

Current state-of-the-art machine learning and neural network algorithms are capable of

50 identifying diagnostic patterns for the purpose of decoding^x a variety of data types, but the
51 inner workings of the resulting model solutions are difficult^x or impossible to interpret.

52 Algorithms that provide such solutions are referred to as *black box* models. Dissections of

53 black box models have been largely uninformative (Zhou, Bau, Oliva, & Torralba, 2019),

54 limiting the potential for researchers to apply the mechanisms underlying successful

55 classification of the data. Still, black box models provide a powerful solution for

56 technological applications such as human-computer interfaces (HCI; for a review, see

57 Lukander, Toivanen, & Puolamäki, 2017). While the internal operations of the model

58 solutions used for HCI applications do not necessarily need to be interpretable to serve their

59 purpose, Lukander et al. (2017) pointed out that the inability to interpret the mechanisms

60 underlying the function of black box solutions impedes the generalizability of these methods,

61 and increases the difficulty of expanding these findings to real life applications. To ground

62 these solutions, researchers guide decoding efforts by using eye movement data and/or

63 models with built-in theoretical assumptions. For instance, eye movement data is processed

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

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

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

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

118 (2014) categorized the same eye movement features used by Greene et al. (2012) with respect
119 to the relative contribution of latent visual or linguistic components of three tasks (visual
120 search, name the picture, name objects in the picture) with 84% accuracy (chance = 33%).
121 While this manipulation is reminiscent of other experiments relying on the bottom-up
122 influence of words and pictures (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016) the eye
123 movements in the Coco and Keller (2014) tasks can be attributed to the occurrence of
124 top-down attentional processes. A conceptually similar follow-up to this study classified
125 tasks along two spatial and semantic dimensions, resulting in 51% classification accuracy
126 (chance = 25%; Król & Król, 2018). A closer look at these results showed that the categories
127 within the semantic dimension were consistently misclassified, suggesting that this level of
128 distinction may require a richer dataset, or a more powerful decoding algorithm. Altogether,
129 there is no measurable index of relative top-down or bottom-up influence, but this body of
130 literature suggests that the relative influence of top-down and bottom-up attentional
131 processes may have a role in determining the decodability of the eye movement data.

132 As shown in Table 1, when eye movement data are prepared for classification, fixation
133 and saccade statistics are typically aggregated along spatial or temporal dimensions,
134 resulting in variables such as fixation density or saccade amplitude (Castelhano et al., 2009;
135 MacInnes et al., 2018; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). The
136 implementation of these statistical methods is meant to explicitly provide the decoding
137 algorithm with characteristics of the eye movement data that are representative of
138 theoretically relevant cognitive processes. For example, MacInnes et al. (2018) attempted to
139 provide an algorithm with data designed to be representative of inputs to the frontal eye
140 fields. In some instances, such as the case of Król and Król (2018), grounding the data using
141 theoretically driven aggregation methods may require sacrificing granularity in the dataset.
142 This means that aggregating the data has the potential to wash out certain fine-grained
143 distinctions that could otherwise be detected. Data structures of any kind can only be
144 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%) |

¹⁴⁵ categories. Given that the cognitive processes underlying distinct tasks are often overlapping¹⁴⁶ (Coco & Keller, 2014), decreasing the granularity of the data may actually limit the potential

¹⁴⁷ of the algorithm to make fine-grained distinctions between diagnostic components underlying
¹⁴⁸ the tasks to be decoded.

¹⁴⁹ The current state of the literature does not provide any firm guidelines for determining
¹⁵⁰ what eye movement features are most meaningful, or what model architectures are best
¹⁵¹ suited for determining mental state from eye movements. The examples provided in Table 1
¹⁵² used a variety of eye movement features and model architectures, most of which were
¹⁵³ effective to some extent. A proper comparison of these outcomes is difficult because these
¹⁵⁴ datasets vary in levels of chance and data quality. Datasets with more tasks to be classified
¹⁵⁵ have lower levels of chance, lowering the threshold for successful classification. Additionally,
¹⁵⁶ datasets with a lower signal-to-noise ratio will have a lower achievable classification accuracy.
¹⁵⁷ For these reasons, outside of re-analyzing the same datasets, there is no consensus on how to
¹⁵⁸ establish direct comparisons of these model architectures. Given the inability to directly
¹⁵⁹ compare the relative effectiveness of the various theoretical approaches present in the
¹⁶⁰ literature, the current study addressed the inverse Yarbus problem by allowing a black box
¹⁶¹ model to self-determine the most informative features from minimally processed eye
¹⁶² movement data.

¹⁶³ The current study explored pragmatic solutions to the problem of classifying task from
¹⁶⁴ eye movement data by submitting unprocessed x-coordinate, y-coordinate, and pupil size
¹⁶⁵ data to a convolutional neural network (CNN) model. Instead of transforming the data into
¹⁶⁶ theoretically defined units, we allowed the network to learn meaningful patterns in the data
¹⁶⁷ on its own. CNNs have a natural propensity to develop low-level feature detectors similar to
¹⁶⁸ the primary visual cortex (e.g., Seeliger et al., 2018); for this reason, they are commonly
¹⁶⁹ implemented for image classification. To test the possibility that the image data are better
¹⁷⁰ suited to the CNN classifier, the data were also transformed from raw timelines into simple
¹⁷¹ image representations. To our knowledge, no study has attempted to address the inverse
¹⁷² Yarbus problem using any combination of the following methods: (1) Non-aggregated data,

¹⁷³ (2) image data format, and (3) a black-box CNN architecture. Given that CNN architectures
¹⁷⁴ are capable of learning features represented in raw data formats, and are well-suited to
¹⁷⁵ decoding multidimensional data that have a distinct spatial or temporal structure, we
¹⁷⁶ expected that a non-theoretically-constrained CNN architecture could be capable of decoding
¹⁷⁷ data at levels consistent with the current state of the art. Furthermore, despite evidence that
¹⁷⁸ black box approaches to the inverse Yarbus problem can impede generalizability (Lukander
¹⁷⁹ et al., 2017), we expected that when testing the approach on an entirely separate dataset,
¹⁸⁰ providing the model with minimally processed data and the flexibility to identify the unique
¹⁸¹ features within each dataset would result in the replication of our initial findings.

¹⁸²

Methods

¹⁸³ Participants

¹⁸⁴ Two separate datasets were used to develop and test the deep CNN architecture. The
¹⁸⁵ two datasets were collected from two separate experiments, which we refer to as Exploratory
¹⁸⁶ and Confirmatory. The participants for both datasets consisted of college students
¹⁸⁷ (Exploratory $N = 124$; Confirmatory $N = 77$) from the University of Nebraska-Lincoln who
¹⁸⁸ participated in exchange for class credit. Participants who took part in the Exploratory
¹⁸⁹ experiment did not participate in the Confirmatory experiment. All procedures and
¹⁹⁰ materials were approved by the University of Nebraska-Lincoln Institutional Review Board
¹⁹¹ prior to data collection.

¹⁹² Materials and Procedures

¹⁹³ Each participant viewed a series of indoor and outdoor scene images while carrying out
¹⁹⁴ a search, memorization, or rating task. For the search task, participants were instructed to
¹⁹⁵ find a small “Z” or “N” embedded in the image. Trials containing a target ($n = 5$) were not
¹⁹⁶ analyzed but were included in the experiment design to encourage searching behavior on
¹⁹⁷ other Search trials. Trials containing the target were excluded because search behavior was

198 likely to stop if the target was found, adding considerable noise to the eye movement data.
199 For consistency between trial types, participants were prompted to indicate if they found a
200 “Z” or “N” at the end each Search trial. For the memorization task, participants were
201 instructed to memorize the image for a forced choice recognition test. At the end of each
202 Memorize trial, the participants were prompted to indicate which of two images was just
203 presented. For the rating task, participants were asked to think about how they would rate
204 the image on a scale from 1 (very unpleasant) to 7 (very pleasant). At the end of the trial,
205 the participants were prompted to provide a rating immediately after viewing the image.
206 The same materials were used in both experiments with a minor variation in the procedures.
207 In the Confirmatory experiment, participants were directed as to where search targets might
208 appear in the image (e.g., on flat surfaces). No such instructions were provided in the
209 Exploratory experiment.

210 In both experiments, participants completed three mixed or uniform blocks of 40 trials
211 ($n = 120$ trials). When the blocks were mixed, the trial types were randomly intermixed
212 within the block. For uniform blocks, each block consisted entirely of one of the three
213 conditions (Search, Memorize, Rate) presented in random order. Each stimulus image was
214 presented for 8 seconds. The pictures were presented in color, with a size of 1024 x 768
215 pixels, subtending a visual angle of 23.8° x 18.0°.

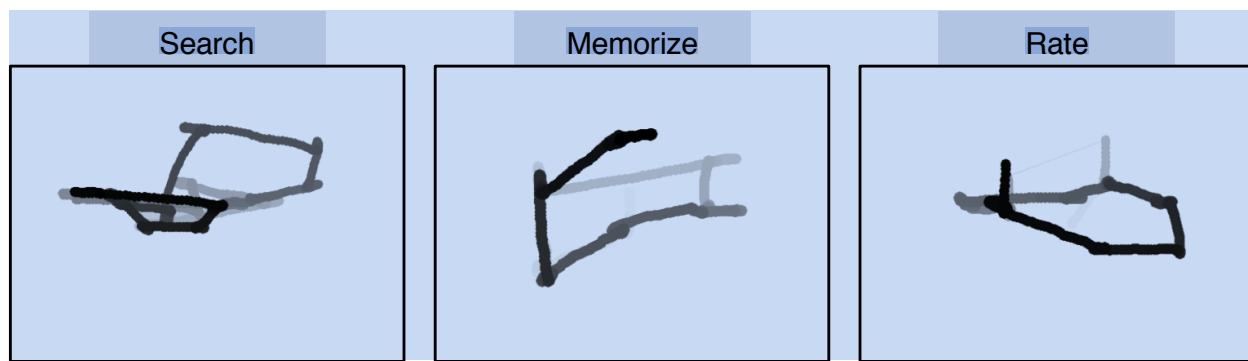
216 Eye movements were recorded using an SR Research EyeLink 1000 eye tracker with a
217 sampling rate of 1000Hz. Only the right eye was recorded. The system was calibrated using
218 a nine-point accuracy and validity test. Errors greater than 1° or averaging greater than 0.5°
219 in total were re-calibrated.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

220 Datasets

221 On some trials, a probe was presented on the screen six seconds after the onset of the
222 trial. To avoid confounds resulting from the probe, only the first six seconds of the data for

²²³ each trial was analyzed. Trials that contained fewer than 6000 samples within the first six
²²⁴ seconds of the trial were excluded before analysis. For both datasets, the trials were pooled
²²⁵ across participants. After excluding trials, the Exploratory dataset consisted of 12,177 of the
²²⁶ 16,740 total trials, and the Confirmatory dataset consisted of 9,301 of the 10,395 total trials.

²²⁷ The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling
²²⁸ time point in the trial were used as inputs to the deep learning classifier. These data were
²²⁹ also used to develop plot image datasets that were classified separately from the raw timeline
²³⁰ datasets. For the plot image datasets, the timeline data for each trial were converted into
²³¹ scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data
²³² point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location
²³³ of the dot, pupil size was used to determine the relative size of the dot, and shading of the
²³⁴ dot was used to indicate the time-course of the eye movements throughout the trial. The
²³⁵ background of the plot images and first data point were white. Each subsequent data point
²³⁶ was one shade darker than the previous data point until the final data point was reached.
²³⁷ The final data point was black. For standardization, pupil size was divided by 10, and one
²³⁸ unit was added. The plots were sized to match the dimensions of the data collection monitor
²³⁹ (1024 x 768 pixels) and then shrunk to (240 x 180 pixels) in an effort to reduce the
²⁴⁰ dimensionality of the data.



²²³ *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.

²⁴¹ **Data Subsets.** The full timeline dataset was structured into three columns

²⁴² representing the x- and y- coordinates, and pupil size for each data point collected in the
²⁴³ first six seconds of each trial. To systematically assess the predictive value of each XYP (i.e.,
²⁴⁴ x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image
²⁴⁵ datasets were batched into subsets that excluded one of the components (i.e., XYØ, XØP,
²⁴⁶ ØYP), or contained only one of the components (i.e., XØØ, ØYØ, ØØP). For the timeline
²⁴⁷ datasets, this means that the columns to be excluded in each data subset were replaced with
²⁴⁸ zeros. The data were replaced with zeros because removing the columns would change the
²⁴⁹ structure of the data. The same systematic batching process was carried out for the image
²⁵⁰ dataset. See Figure 2 for an example of each of these image data subsets.

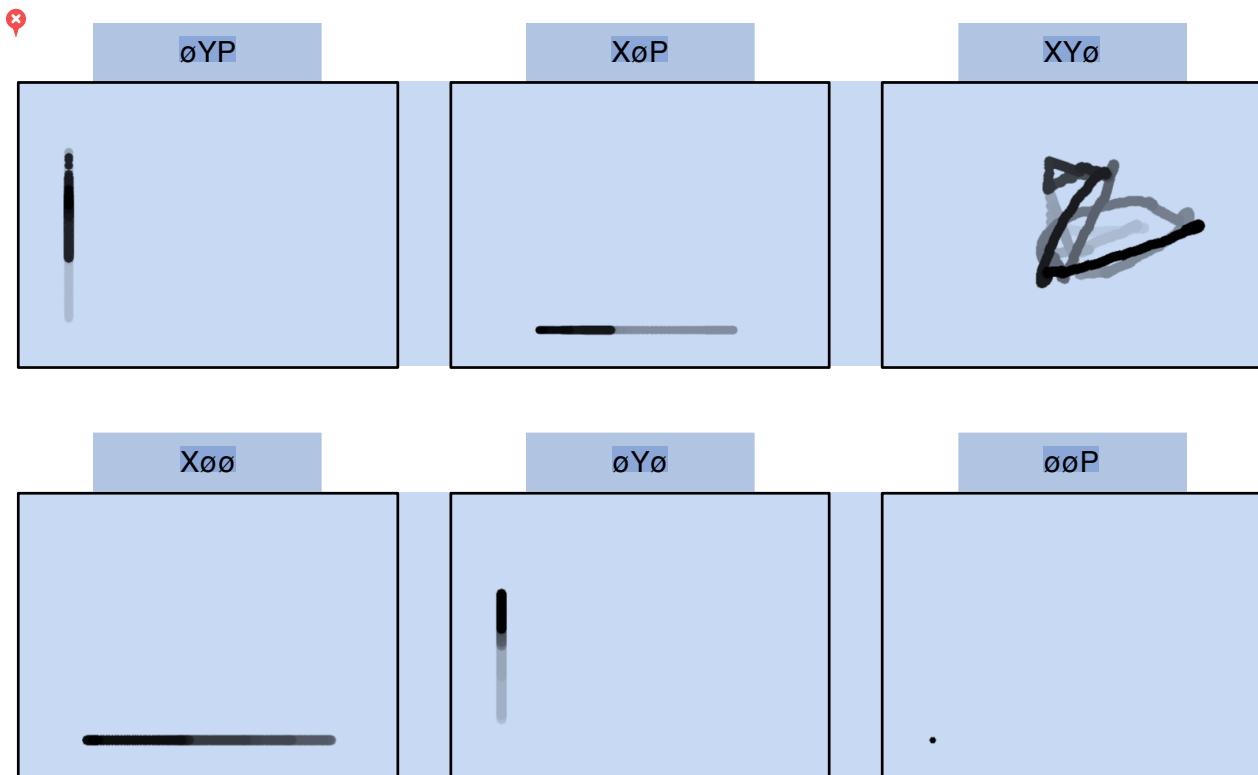


Figure 2. Plot images were used to represent each type of data subset. As with the trials in the full XYP dataset, the time course of the eye movements was represented by the shading of the dot. The first sample of each trial was white, and the last sample was black.

251 Classification

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

258 The development of these models was not guided by any formal theoretical assumptions
259 regarding the patterns or features likely to be extracted by the classifier. Like many HCI
260 models, the development of these models followed general intuitions concerned with building
261 a model architecture capable of transforming the data inputs into an interpretable feature
262 set that would not overfit the dataset. The models were developed using version 0.3b of the
263 DeLINEATE toolbox, which operates over a Keras backend ([https://delineate.it/](https://delineate.it;);
264 Kuntzelman et al., under review). Each training/test iteration randomly split the data so
265 that 70% of the trials were allocated to training, 15% to validation, and 15% to testing.
266 Training of the model was stopped when validation accuracy did not improve over the span
267 of 100 epochs. Once the early stopping threshold was reached, the resulting model was
268 tested on the held-out test data. This process was repeated 10 times for each model,
269 resulting in 10 classification accuracy scores for each model. The resulting accuracy scores
270 were used for the comparisons against chance and other datasets or data subsets.

271 The models were developed and tested on the Exploratory dataset. Model
272 hyperparameters were adjusted until the classification accuracies appeared to peak. The
273 model architecture with the highest classification accuracy on the Exploratory dataset was
274 trained, validated, and tested independently on the Confirmatory dataset. This means that
275 the model that was used to analyze the Confirmatory dataset was not trained on the
276 Exploratory dataset. The model architectures used for the timeline and plot image datasets

277 are shown in Figure 3.

278 **Analysis**

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

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

297 **Results**

298 **Timeline Data Classification**

299 **Exploratory.** Classification accuracies for the XYP timeline dataset were well above
300 chance (chance = .33; $M = .526$, $SD = .018$; $t_{(9)} = 34.565$, $p < .001$). Accuracies for
301 classifications of the batched data subsets were all better than chance (see Figure 4). As

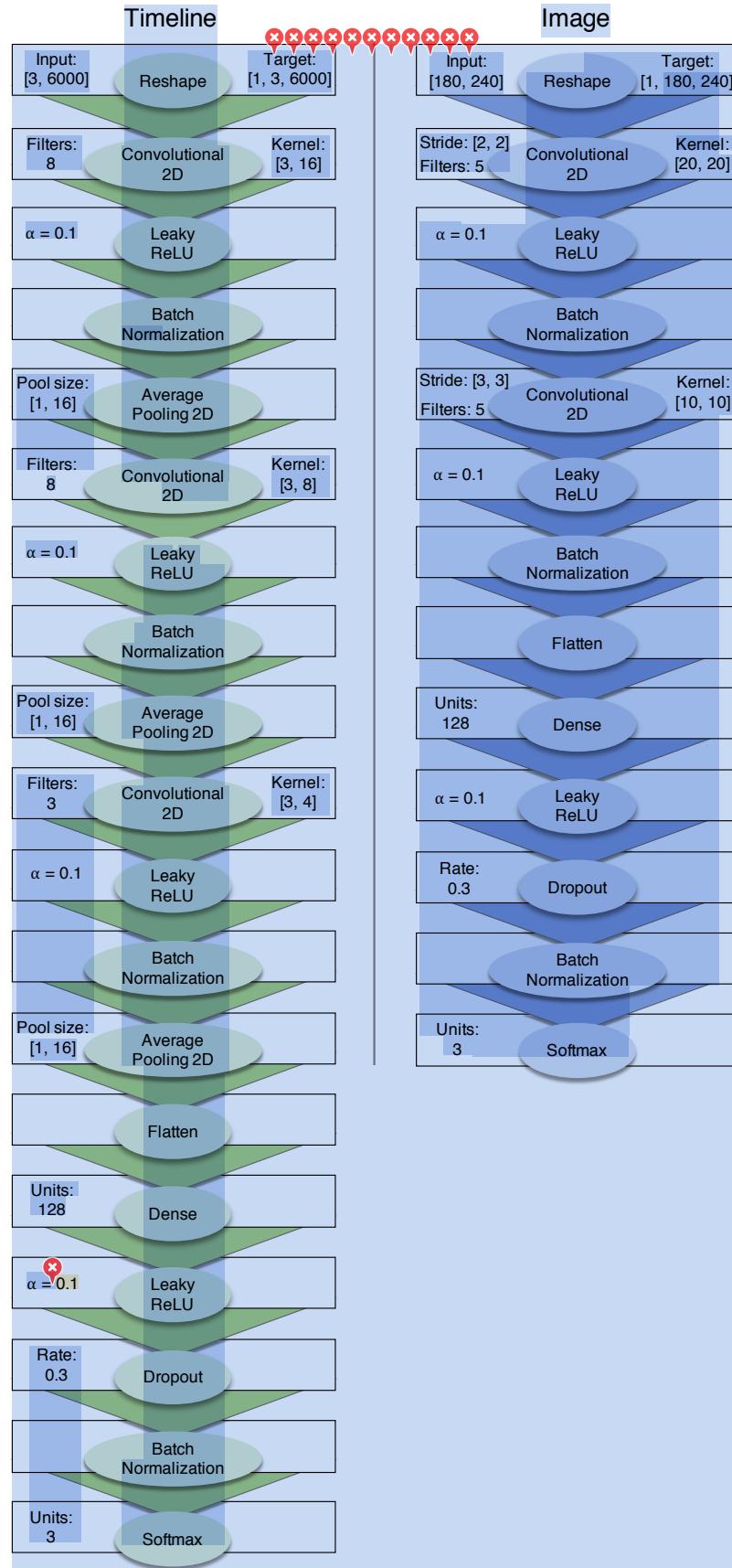


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm.

302 shown in the confusion matrices displayed in Figure 5, the data subsets with lower overall
 303 classification accuracies almost always classified the Memorize condition at or below chance
 304 levels of accuracy. Misclassifications of the Memorize condition were split relatively evenly
 305 between the Search and Rate conditions.

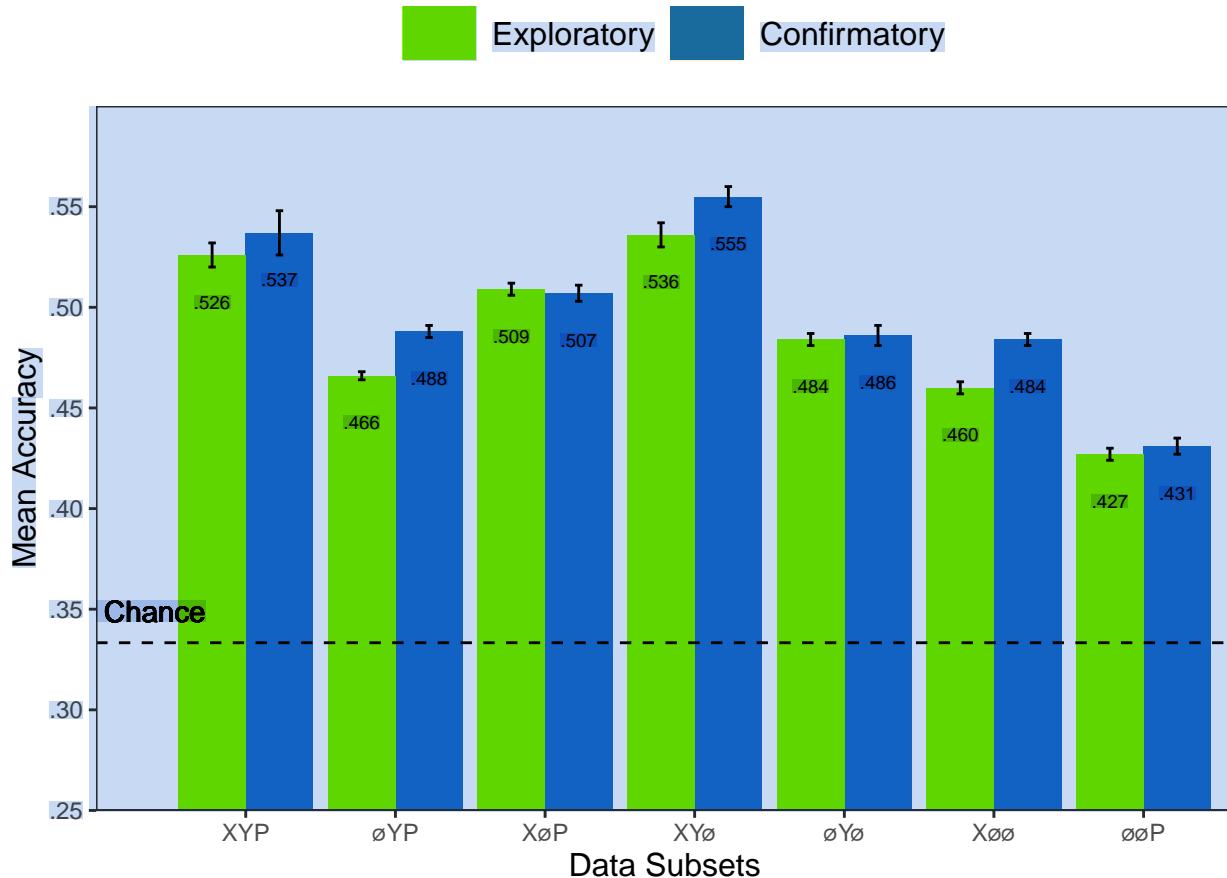


Figure 4. All of the data subsets were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

306 There was a difference in classification accuracy for the XYP dataset and the subsets
 307 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
 308 classification accuracies were not affected by the removal of pupil size or y-coordinate data
 309 (see Table 2). The null effect present when pupil size was removed suggests that the pupil
 310 size data were not contributing unique information that was not otherwise provided by the x-
 311 and y-coordinates. A strict significance threshold of $\alpha = .05$ implies the same conclusion for
 312

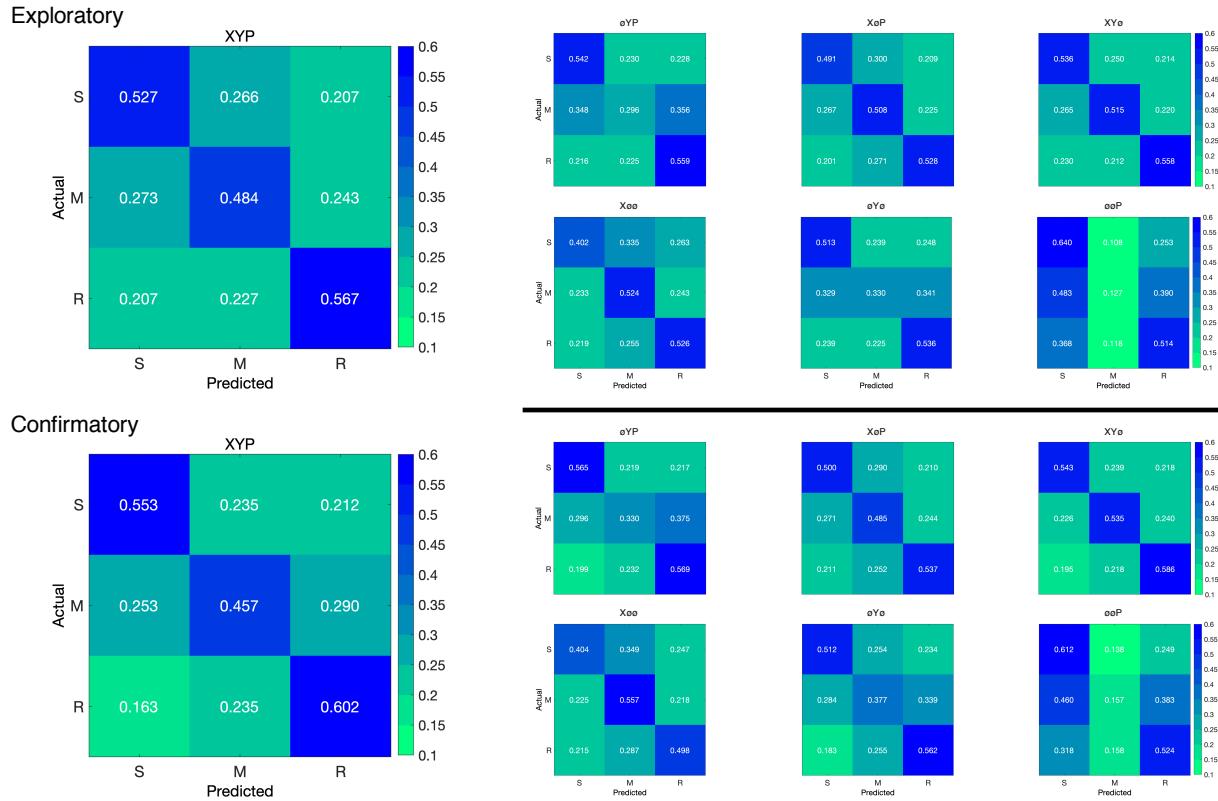


Figure 5. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

313 the y-coordinate data, but the relatively low degrees of freedom ($df = 18$) and the borderline
 314 observed p -value ($p = .056$) afford the possibility that there exists a small effect. However,
 315 classification for the $\emptyset Y P$ subset was significantly lower than the XYP dataset, showing that
 316 the x-coordinate data were uniquely informative to the classification.

Table 2
Timeline Subset Comparisons

| Comparison | Exploratory | | Confirmatory | |
|---|-------------|----------|--------------|----------|
| | <i>t</i> | <i>p</i> | <i>t</i> | <i>p</i> |
| XYP vs. $\emptyset Y P$ | 9.420 | < .001 | 5.210 | < .001 |
| XYP vs. XoP | 2.645 | .056 | 3.165 | .016 |
| XYP vs. XYo | 1.635 | .372 | 1.805 | .288 |
| XoO vs. $\emptyset Y O$ | 5.187 | < .001 | 0.495 | .874 |
| XoO vs. $\emptyset \emptyset P$ | 12.213 | < .001 | 10.178 | < .001 |
| $\emptyset Y O$ vs. $\emptyset \emptyset P$ | 7.026 | < .001 | 9.683 | < .001 |

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

³¹⁸ subsets ($F_{(2,27)} = 75.145, p < .001, \eta^2 = 0.848$). Post-hoc comparisons showed that
³¹⁹ classification accuracy for the $\emptyset\emptyset P$ subset was lower than the $X\emptyset\emptyset$ and $\emptyset Y\emptyset$ subsets.
³²⁰ Classification accuracy for the $X\emptyset\emptyset$ subset was higher than the $\emptyset Y\emptyset$ subset. Altogether,
³²¹ these findings suggest that pupil size data was the least uniquely informative to classification
³²² decisions, while the x-coordinate data was the most uniquely informative.

³²³ **Confirmatory.** Classification accuracies for the Confirmatory XYP timeline dataset
³²⁴ were well above chance ($M = .537, SD = 0.036, t_{(9)} = 17.849, p < .001$). Classification
³²⁵ accuracies for the data subsets were also better than chance (see Figure 4). Overall, there
³²⁶ was high similarity in the pattern of results for the Exploratory and Confirmatory datasets
³²⁷ (see Figure 4). Furthermore, the general trend showing that pupil size was the least
³²⁸ informative eye tracking data component was replicated in the Confirmatory dataset (see
³²⁹ Table 2). Also in concordance with the Exploratory timeline dataset, the confusion matrices
³³⁰ for these data revealed that the Memorize task was mis-classified more often than the Search
³³¹ and Rate tasks (see Figure 5).

³³² To test the generalizability of the model architecture, classification accuracies for the
³³³ XYP Exploratory and Confirmatory timeline datasets were compared. The Shapiro-Wilk
³³⁴ test for normality indicated that the Exploratory ($W = 0.937, p = .524$) and Confirmatory
³³⁵ ($W = 0.884, p = .145$) datasets were normally distributed, but Levene's test indicated that
³³⁶ the variances were not equal, $F_{(1,18)} = 8.783, p = .008$. Welch's unequal variances t -test did
³³⁷ not show a difference between the two datasets, $t_{(13.045)} = 0.907, p = .381$, Cohen's $d =$
³³⁸ 0.406. These findings indicate that the deep learning model decoded the Exploratory and
³³⁹ Confirmatory timeline datasets equally well, but the Confirmatory dataset classifications
³⁴⁰ were less consistent across training/test iterations (as indicated by the increase in standard
³⁴¹ deviation).

✖ 342 Plot Image Classification

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

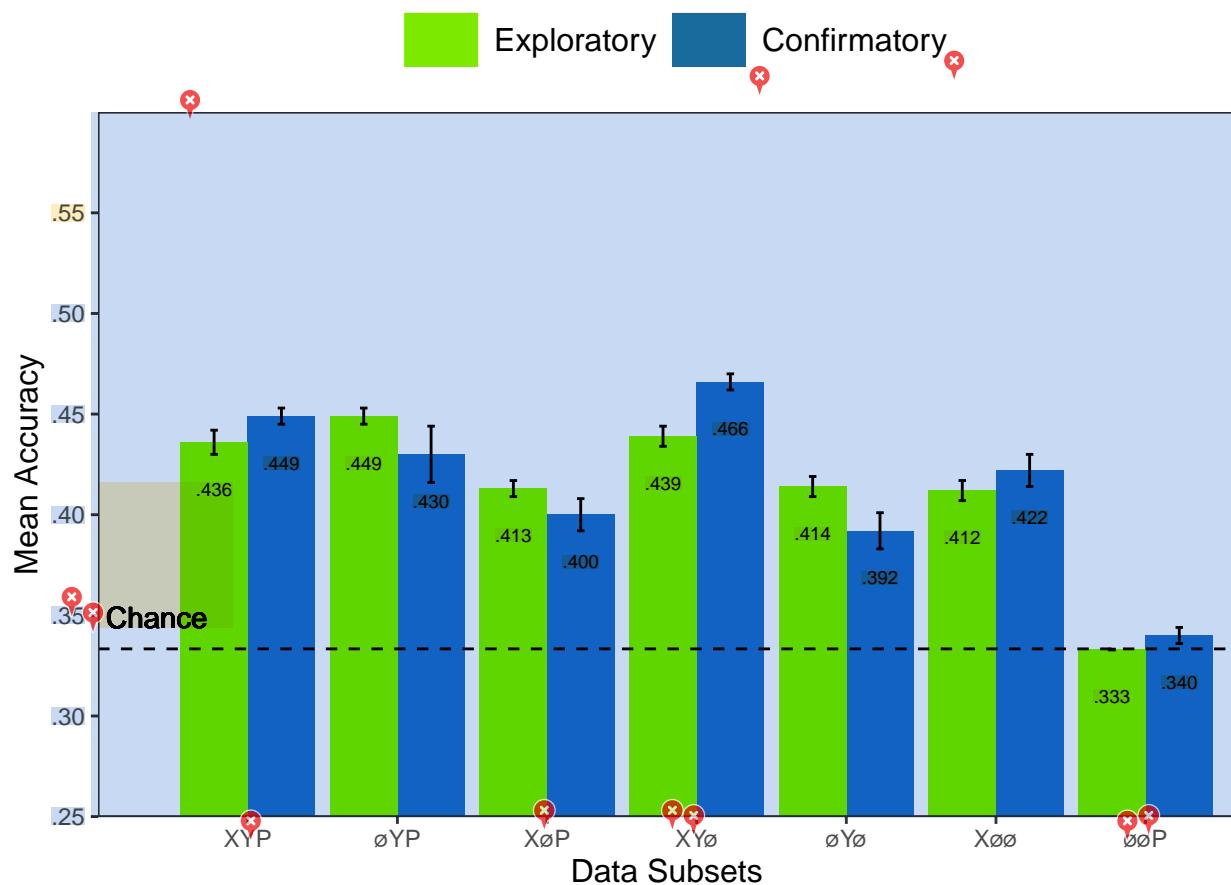


Figure 6. ✖ All of the data subsets except for the Exploratory $\emptyset\emptyset P$ dataset were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

351 There was a difference in classification accuracy between the XYP dataset and the data sets for the other six subsets. There was no difference in classification accuracy between the Exploratory and Confirmatory datasets for the XYP, $\emptyset Y P$, $X \emptyset P$, $X Y \emptyset$, $\emptyset Y \emptyset$, and $X \emptyset \emptyset$ subsets. There was a difference in classification accuracy between the Exploratory and Confirmatory datasets for the $\emptyset \emptyset P$ subset.

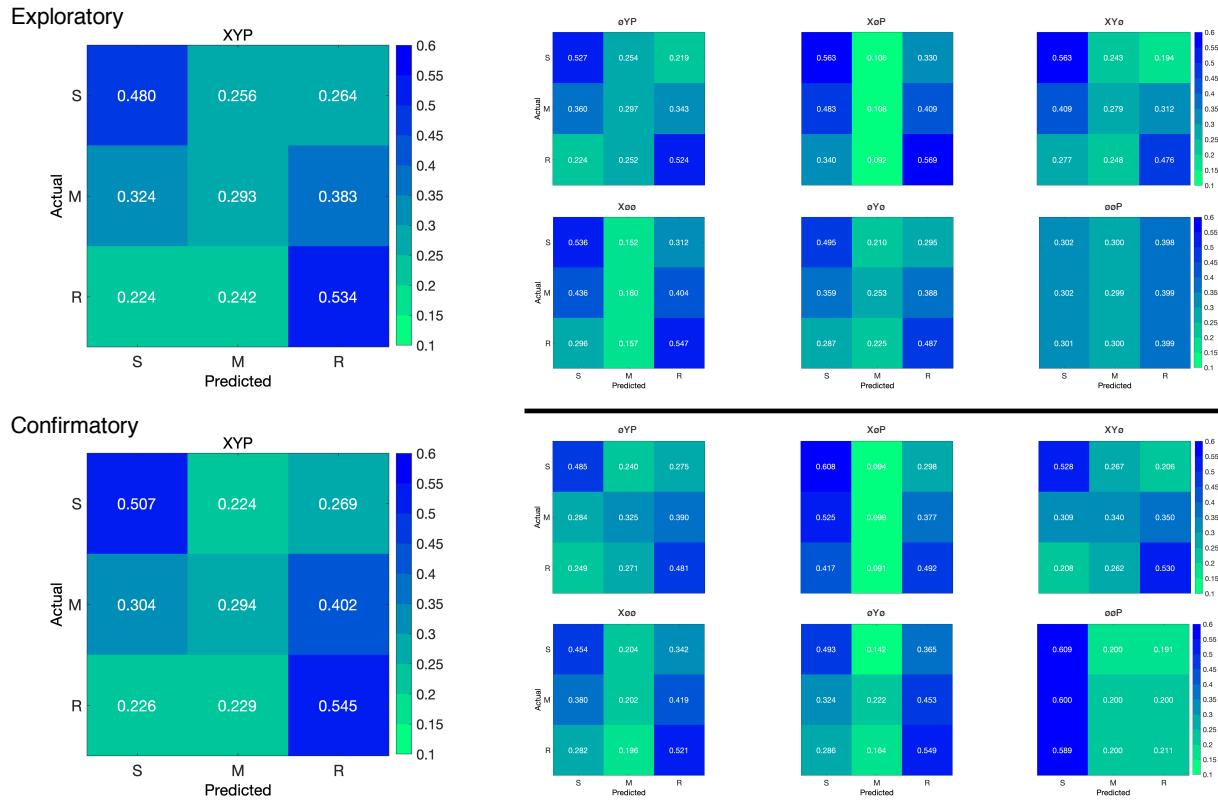


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

352 subsets ($F_{(4,45)} = 7.093, p < .001, \eta^2 = .387$). Post-hoc comparisons showed that compared
 353 to the XYP dataset, there was no effect of removing pupil size or the x-coordinates, but
 354 classification accuracy was worse when the y-coordinates were removed (see Table 3).

Table 3
Image Subset Comparisons

| Comparison | Exploratory | | Confirmatory | |
|--|-------------|--------|--------------|--------|
| | t | p | t | p |
| XYP vs. \emptyset YP | 1.792 | .391 | 1.623 | .491 |
| XYP vs. XoP | 2.939 | .039 | 4.375 | < .001 |
| XYP vs. XYo | 0.474 | .989 | 1.557 | .532 |
| XoO vs. \emptyset Y \emptyset | 0.423 | .906 | 2.807 | .204 |
| XoO vs. \emptyset oP | 13.569 | < .001 | 5.070 | < .001 |
| \emptyset Y \emptyset vs. \emptyset oP | 13.235 | < .001 | 7.877 | < .001 |

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

³⁵⁷ homogeneity of variances: $F_{(2,17.993)} = 228.137, p < .001, \eta^2 = .899$). Post-hoc comparisons
³⁵⁸ showed that there was no difference in classification accuracies for the XØØ and ØYØ
³⁵⁹ subsets, but classification for the ØØP subset were less accurate than the XØØ and ØYØ
³⁶⁰ subsets.

³⁶¹ **Confirmatory.** Classification accuracies for the XYP confirmatory image dataset
³⁶² were well above chance ($M = .449, SD = 0.012, t_{(9)} = 31.061, p < .001$), but were less
³⁶³ accurate than the classifications of the confirmatory timeline dataset ($t_{(18)} = 11.167 p <$
³⁶⁴ .001). Accuracies for classifications of the data subsets were also all better than chance (see
³⁶⁵ Figure 6). The confusion matrices followed the pattern showing that the Memorize condition
³⁶⁶ was confused most often, and was relatively evenly mis-identified as a Search or Rate trial
³⁶⁷ (see Figure 7). As with the timeline data, the general trend showing that pupil size data was
³⁶⁸ the least informative to the model was replicated in the Confirmatory dataset (see Table 3).

³⁶⁹ To test the generalizability of the model architecture, the classification accuracies for
³⁷⁰ the XYP Exploratory and Confirmatory plot image datasets were compared. The
³⁷¹ independent samples *t*-test comparing the classification accuracies for the Exploratory and
³⁷² Confirmatory plot image datasets did not show a significant difference, $t_{(18)} = 1.777, p =$
³⁷³ .092, Cohen's *d* = 0.795.

³⁷⁴

Discussion

³⁷⁵ The present study aimed to produce a practical and reliable example of a black box
³⁷⁶ solution to the inverse Yarbus problem. To implement this solution, we classified raw
³⁷⁷ timeline and minimally processed plot image data using a CNN model architecture. To our
³⁷⁸ knowledge, this study was the first to provide a solution to determining mental state from
³⁷⁹ eye movement data using each of the following: (1) Non-aggregated eye tracking data (i.e.,
³⁸⁰ raw x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see
³⁸¹ Figure 2), and (3) a black box CNN architecture. This study probed the relative predictive
³⁸² value of the x-coordinate, y-coordinate, and pupil size components of the eye movement data

383 using a CNN. The CNN was able to decode the timeline and plot image data better than
384 chance, although only the timeline datasets were decoded with accuracies comparable to
385 other state-of-the-art approaches. Datasets with lower classification accuracies were not able
386 to differentiate the cognitive processes underlying the Memorize task from the cognitive
387 processes underlying the Search and Rate tasks. Decoding subsets of the data revealed that
388 pupil size was the least uniquely informative component of the eye movement data. This
389 pattern of findings was consistent between the Exploratory and Confirmatory datasets.

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

404 When considering the superior performance of the timeline data (vs., plot image data),
405 we must also consider the differences in the model architectures. Because the structures of
406 the timeline and plot image data formats were different, the models decoding those data
407 structures also needed to be different. Both model architectures were optimized individually
408 on the Exploratory dataset before being tested on the Confirmatory dataset. For both 

409 timeline and plot image formats, there was good replicability between the Exploratory and
410 Confirmatory datasets, demonstrating that these architectures performed similarly from
411 experiment to experiment. An appropriately tuned CNN should be capable of learning any
412 arbitrary function, but given that the upper bound for decodability of these datasets is
413 unknown, there is the possibility that a model architecture exists that is capable of
414 classifying the plot image data format more accurately than the model used to classify the
415 timeline data. Despite this possibility, the convergence of these findings with other studies
416 (see Table 1) suggests that the results of this study are approaching a ceiling for the
417 potential to solve the inverse Yarbus problem with eye movement data. Although the true
418 capacity to predict mental state from eye movement data is unknown, standardizing datasets
419 in the future could provide a point for comparison that can more effectively indicate which
420 methods are most effective at solving the inverse Yarbus problem.

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

431 To more clearly understand how the different tasks influenced the decodability of the
432 eye movement data, additional analyses were conducted on the Exploratory and
433 Confirmatory timeline datasets (see Appendix). For the main supplementary analysis, the
434 data subsets were submitted to the model in 2-category task sets. In addition to the

435 supplementary analysis, the results from the primary analysis were re-calculated from
436 3-category task sets to 2-category task sets. These analyses showed a tendency for the model
437 to mis-classify the Search and Rate trials as Memorize. In the primary analyses, the
438 Memorize condition was predicted with the lowest accuracy, but mis-classifications of the
439 Search and Rate trials were most often categorized as Memorize. As a whole, this pattern of
440 results indicated a general bias for uncertain trials to be categorized as Memorize. Overall,
441 the findings from this supplemental analysis show that conclusions drawn from comparisons
442 between approaches that do not use the same task sets, or the same number of tasks could
443 be potentially uninterpretable because the features underlying the task categories are
444 interpreted differently by the neural network algorithm.

445 When determining the relative contributions of the eye movement features used in
446 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently the
447 least uniquely informative. When pupil size was removed from the Exploratory and
448 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,
449 XYP dataset). Furthermore, classification accuracy of the $\emptyset\emptyset P$ subset was the lowest of all
450 of the data subsets, and in one instance, was no better than chance. Although these findings
451 indicate that, in this case, pupil size was a relatively uninformative component of the eye
452 movement data, previous research has associated changes in pupil size as indicators of
453 working memory load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004),
454 arousal (Wang et al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The
455 results of the current study indicate that the changes in pupil size associated with these
456 underlying processes were not useful in delineating the tasks being classified (i.e., Search,
457 Memorize, Rate), potentially because these tasks did not evoke a reliable pattern of changes
458 in pupil size. Additionally, properties of the stimuli known to influence pupil size, such as
459 luminance and contrast, were not controlled in these datasets. Given that stimuli were
460 randomly assigned, there is the potential that uncontrolled stimulus properties known to
461 affect pupil size made impeded the CNN's capacity to detect patterns in the pupil size data.

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

Classifying mental state from eye movement data is often carried out in an effort to advance technology to improve educational outcomes, strengthen the independence of physically and mentally handicapped individuals, or improve HCI's (Koochaki & Najafizadeh, 2018). Given the previous questions raised regarding the reliability and generalizability of black-box CNN classification, the current study first tested models on an exploratory dataset, then confirmed the outcome using a second independent dataset. Overall, the findings of this study indicate that this black-box approach is capable of producing a stable and generalizable outcome. Additionally, the supplementary analyses showed that different task sets, or a different number of tasks, could lead the algorithm to

489 interpret features differently, which should be taken into account when comparing task
490 classification approaches. Future studies that incorporate features from the stimulus might
491 have the potential to surpass current state-of-the-art classification. According to Bulling,
492 Weichel, and Gellersen (2013), incorporating stimulus feature information into the dataset
493 may provide improve accuracy relative to decoding gaze location data and pupil size.
494 Alternatively, Borji and Itti (2014) suggested that accounting for salient features in the the
495 stimulus might leave little to no room for theoretically defined classifiers to consider mental
496 state. Future research should examine the potential for the inclusion of stimulus feature
497 information in addition to the eye movement data to boost black-box CNN classification
498 accuracy of image data beyond that of timeline data.

499

References

- 500 Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the
501 importance of spatial distribution, dynamics, and image features. *Neurocomputing*,
502 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- 503 Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task.
504 *Journal of Vision*, 14(3), 29–29. <https://doi.org/10.1167/14.3.29>
- 505 Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level
506 contextual cues from human visual behaviour. In *Proceedings of the SIGCHI
507 Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris,
508 France: ACM Press. <https://doi.org/10.1145/2470654.2470697>
- 509 Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye
510 movement control during active scene perception. *Journal of Vision*, 9(3), 6–6.
511 <https://doi.org/10.1167/9.3.6>
- 512 Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using
513 eye-movement features. *Journal of Vision*, 14(3), 11–11.
514 <https://doi.org/10.1167/14.3.11>
- 515 DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited.
516 *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- 517 Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict
518 observers' task from eye movement patterns. *Vision Res*, 62, 1–8.
519 <https://doi.org/10.1016/j.visres.2012.03.019>
- 520 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers'
521 task from eye movement patterns. *Vision Research*, 103, 127–142.

- 522 <https://doi.org/10.1016/j.visres.2014.08.014>
- 523 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).
524 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937.
525 <https://doi.org/10.1371/journal.pone.0064937>
- 526 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,
527 154(3756), 1583–1585. Retrieved from <https://www.jstor.org/stable/1720478>
- 528 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting
529 an observer's task using multi-fixation pattern analysis. In *Proceedings of the*
530 *Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290).
531 Safety Harbor, Florida: ACM Press. <https://doi.org/10.1145/2578153.2578208>
- 532 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the
533 dual-task paradigm as measured through behavioral and psychophysiological
534 responses. *Psychophysiology*, 41(2), 175–185.
535 <https://doi.org/10.1111/j.1469-8986.2004.00147.x>
- 536 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze Patterns.
537 In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).
538 <https://doi.org/10.1109/BIOCAS.2018.8584665>
- 539 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive processes
540 involved in the task from spatial eye-movement patterns. *Psychological Research*.
541 <https://doi.org/10.1007/s00426-018-0996-5>
- 542 Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from Gaze
543 in Naturalistic Behavior: A Review. *International Journal of Mobile Human-Computer Interaction*, 9(4), 41–57. <https://doi.org/10.4018/IJMHCI.2017100104>

- 545 MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative
546 Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*,
547 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- 548 Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011).
549 Examining the influence of task set on eye movements and fixations. *Journal of*
550 *Vision*, 11(8), 17–17. <https://doi.org/10.1167/11.8.17>
- 551 Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and
552 counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*
553 (2006), 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- 554 Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., &
555 van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and
556 decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266.
557 <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- 558 Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus,
559 Eye Movements, and Vision. *I-Perception*, 1(1), 7–27. <https://doi.org/10.1068/i0382>
- 560 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018).
561 Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional
562 Face Task. *Frontiers in Neurology*, 9. <https://doi.org/10.3389/fneur.2018.01029>
- 563 Yarbus, A. (1967). Eye Movements and Vision. Retrieved January 24, 2019, from
564 [http://wexler.free.fr/library/files/yarbus%20\(1967\)%20eye%20movements%20and%20vision.pdf](http://wexler.free.fr/library/files/yarbus%20(1967)%20eye%20movements%20and%20vision.pdf)
- 566 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep
567 Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K.
568 Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and*

569 *Visualizing Deep Learning* (pp. 243–252). Cham: Springer International Publishing.

570 https://doi.org/10.1007/978-3-030-28954-6_12

Appendix

571 Additional analyses were conducted in an attempt to clarify the effect of task on classification
 572 accuracy. These supplementary analyses were not seen as central to the current study, but
 573 could prove^x to be informative to researchers attempting to replicate or extend these findings
 574 in the future. The results from the primary analysis showed that classification accuracies
 575 were the lowest for the Memorize condition. To further understand why classification
 576 accuracy was lower for the Memorize condition than it was for the Search or Rate condition,
 577 the Exploratory and Confirmatory timeline datasets were systematically batched into subsets^x
 578 with the Search (S), Memorize (M), or Rate (R) condition removed (i.e., $\emptyset\text{MR}$, $S\emptyset R$, $SM\emptyset$).

579 All of the data subsets analyzed in this supplementary analysis were decoded with
 580 better than chance accuracy (see Figure A1a). The same pattern of results was observed in
 581 both the Exploratory and Confirmatory datasets. When the Memorize condition was
 582 removed, classification accuracy improved (see Table A1; see Figure A1a). When the Rate
 583 condition was removed, classification was the worst. When the Memorize condition was
 584 included (i.e., $SM\emptyset$ and $\emptyset\text{MR}$), mis-classifications were biased toward Memorize, and the
 585 Memorize condition was more accurately predicted than the Search and Rate conditions (see
 586 Figure A2).

Table A1
Supplementary Subset Comparisons

| Comparison | Exploratory | | Confirmatory | |
|---|-------------|----------|--------------|----------|
| | <i>t</i> | <i>p</i> | <i>t</i> | <i>p</i> |
| $\emptyset\text{MR}$ vs. $S\emptyset R$ | 3.248 | .008 | 3.094 | .012 |
| $\emptyset\text{MR}$ vs. $SM\emptyset$ | 2.875 | .021 | 2.923 | .018 |
| $S\emptyset R$ vs. $SM\emptyset$ | 6.123 | < .001 | 6.017 | < .001 |

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

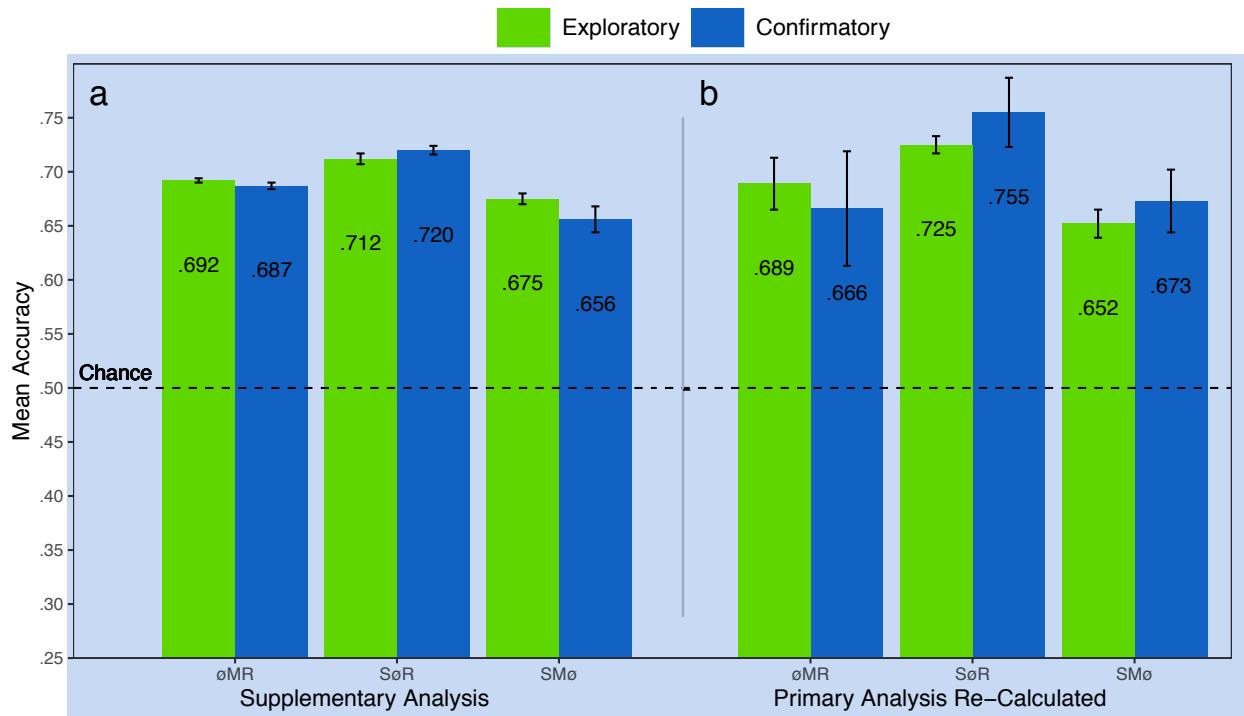


Figure A1. The graph represents the average accuracy reported for each subset of the Exploratory and Confirmatory timeline data for (a) the re-calculated accuracies from the primary analysis, and (b) the supplementary analysis. All of the data subsets were decoded at levels better than chance (.50). 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 (<https://delineate.it>; Kuntzelman et al., under review) guaranteed an equal number of trials in the test set are 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 SØR would be calculated with the following: $\text{Prediction}_{(S,S,S \otimes R)} = \text{Prediction}_{(S,S,S \otimes R)} / (\text{Prediction}_{(S,S,S \otimes)} + \text{Prediction}_{(S,R,S \otimes R)})$, where $\text{Prediction}_{(S,R,S \otimes R)}$ is

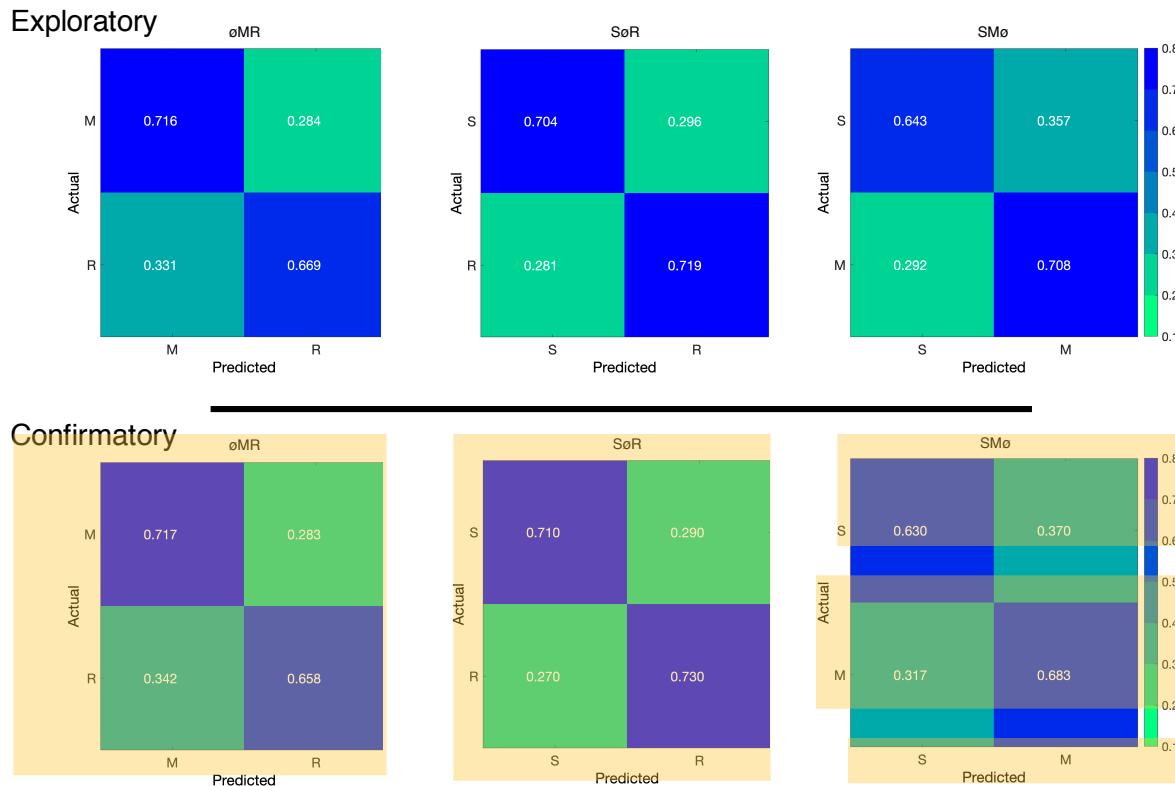


Figure A2. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

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

604 The results for the re-calculated predictions followed a pattern similar to the

605 supplementary analysis (see Figure A1b). This is supported by the persisting tendency of

606 the algorithm to mis-classify Search and Rate trials in the SMø and øMR subsets as

607 Memorize (see Figure A3). Looking back at the primary analysis, the 3-category

608 classifications predicted the Memorize conditions with the lowest accuracy (c.f., Search and

609 Rate conditions), mis-classifications of the Search and Rate conditions were most often

610 categorized as Memorize (see Figure 5). This overall pattern points toward a general bias to

611 categorize uncertain trials as Memorize.

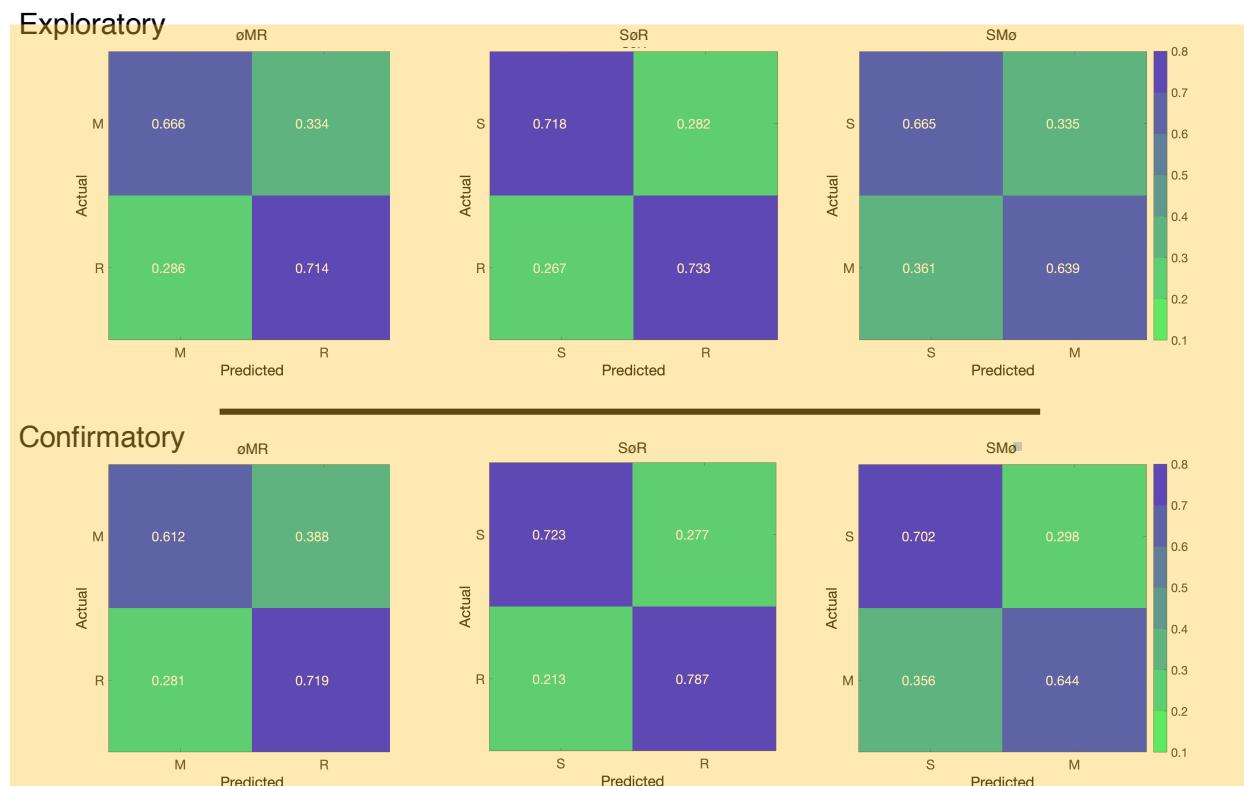


Figure A3. The confusion matrices represent a re-calculation of the classification accuracies for each category from the primary analysis. This re-calculation is meant to make the accuracies presented in the primary analysis (chance = .33) equivalent to the classification accuracies presented in the supplementary analysis (chance = .50).