We would like to start by thanking the editor and reviewers who took the time to carefully review and evaluate our manuscript “Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements” originally submitted to *Journal of Vision* on September 14, 2020. Overall, we feel that the changes we have implemented in response to the thoughtful comments and recommendations from the editor and reviewers have improved the quality of the manuscript.

The enclosed document provides a verbatim copy of the initial review comments and our responses recorded in-line. We made the effort to respond completely and concisely to each comment, in addition to providing a short description of how the comment or recommendation is addressed in the re-submitted manuscript.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Manuscript ID JOV-07646-2020 titled "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements"  
  
Dear Mr. Cole:  
  
Reviews have been returned for your paper, and several points have been raised which will need to be addressed before a recommendation can be made.  
  
Both reviewers make valuable comments, particularly reviewer #2 has very detailed and insightful suggestions. Please pay careful attention to each of the points raised before submitting your revision. It is likely that your revised manuscript will be returned to at least one of the previous referees. Sometimes, an expert who was not part of the initial review process will also be invited to comment on the revision. Criticisms that were not mentioned during the initial review may arise at a future stage of the peer review process.  
  
Reviewers may make recommendations as to the suitability of a paper for publication in JOV, but these recommendations do not guarantee eventual acceptance.  
  
Please prepare a point-by-point response to the suggestions of the reviewers. This can be a Word or PDF file to be uploaded with the rest of the manuscript files under the "Author Response to Reviewer(s)" file type. Please be as specific as possible in explaining the changes made to your manuscript.  
  
If we do not receive a revision within 90 days, we will consider the manuscript withdrawn from the Journal of Vision.  
  
Refer to the instructions at the end of this email on how to submit your revision. Please contact Kiyah Morrison at the JOV Editorial Office at [kmorrison@arvo.org](mailto:kmorrison@arvo.org) or 240-221-2933 if you have any questions.  
  
Thank you for giving us the opportunity to review your paper. I look forward to receiving your revised work.  
  
Sincerely,  
Felix Wichmann  
Editorial Board Member  
JOV  
  
Editor, Journal of Vision  
  
INSTRUCTIONS  
You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text.  
  
Please prepare a point-by-point response to the suggestions of the reviewers. This can be a Word or PDF file to be uploaded with the rest of the manuscript files. Please be as specific as possible in explaining the changes made to your manuscript.  
  
IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Editor Comments:  
  
  
  
Reviewer #1 (Comments for the Author (Required)):  
  
Review on Cole et al. "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements"   
  
Topic:   
The authors trained a CCN classifier for task classification from two eye-movement datasets. The key innovative point is that time-dependent "raw" data of eye positions were used. The authors demonstrate a reliable black box solution to task classification from eye movements.  
  
Comments:  
Greene et al. and Borji & Itti investigated the long-standing Yarbus problem before. The current work, the authors claim to have contributed to its solution. The fundamental problem, from my perspective, is related to the definition of the Yarbus challenge. As Greene et al. put it, "Yarbus argued that changing the information that an observer is asked to obtain from an image drastically changes his pattern of eye movements." The "pattern" is clearly meant to be related to series fixations and saccades, not raw data. Therefore, the relation between the current study and the Yarbus problem is rather intransparent and complicated. It is known that fixational eye movements and microsaccades are highly viewer-specific and can be strongly modulated by task. Therefore, it could well be that the current manuscript's results are completely based on aspects of the data that are in a rather indirect connection to the original Yarbus problem.   
Two possible solutions seem viable: First, compare results on the raw eye traces with processes data of saccade and fixation sequences. Second, rewrite the manuscript without the attempt to solve the Yarbus problem. The latter solution would make a good contribution. The first is necessary, if the authors that their black box algorithm is classifying "mental state" from eye movement data without discussing the potential bypass from cognition to fixational eye movements to classification, which has little to do with the intended long-standing Yarbus challenge.  
EOF   
  
Reviewer #1 has provided a compact statement describing a perceived theoretical disconnect between our work and the current literature. We have chosen to break down our response to this comment in sections below:

(a) The fundamental problem with this manuscript, in the eyes of Reviewer #1, is in regard to the definition of the “Yarbus challenge.” This problem appears to stem from a quote the reviewer provides from a manuscript by Greene et al. (2012), in which the reviewer asserts that the word “pattern” is “clearly meant to be related to series of fixations and saccades, not raw data.” For this reason, the reviewer believes that our work is not clearly related to the “Yarbus problem” and that the results presented in our manuscript could be based on aspects of the data not relevant to the “original Yarbus problem.”

We felt that it was important to address specific assertions made by the reviewer separate from the proposed solutions the reviewer provided. Here, we discuss important distinctions between specific points made by the reviewers and what we have discovered in the literature. We believe that the apparent fundamental disconnect between our approach and the other approaches to the inverse Yarbus problem is mainly due to some misunderstandings that we have attempted to address in the manuscript by clarifying language used throughout the manuscript. The specific comments we were able to address are presented below:

1. Yarbus challenge/problem

The main purpose of the paper is to categorize task-at-hand from minimally processed eye movement data, also known as the inverse Yarbus process (Haji-Abolhassani & Clark, 2014). The original Yarbus problem (Yarbus, 1967), was to understand how the task influences eye movement data, a definition that follows with the Greene et al. quote presented by the reviewer. This means that the Yarbus problem, and the inverse Yarbus problem are two separate (albeit related) problems. From this comment, we cannot discern whether the reviewer is aware of this distinction, if the reviewer characterized our findings with the Yarbus problem because this is how our manuscript read, or if they just chose to use the terms “Yarbus challenge” and “Yarbus problem” out of convenience (less to type).

Regardless of the source for this apparent misunderstanding, we feel that this is a very important distinction to be made. In order to avoid future misunderstanding, we made it a point to address this issue by further clarifying the distinction between the “Yarbus process” and the “inverse Yarbus process” within the manuscript where appropriate.

All of the following comments in response to the comments from Reviewer #1 treat references from the reviewer to the “Yarbus Challenge” or the “Yarbus problem” as meaning to refer to the “inverse Yarbus problem.”

1. Borji & Itti

The reviewer specifically points to Greene et al. (2012) and Borji & Itti (2014) as examples of studies of the “Yarbus problem.” These studies are supposed to present what the reviewer perceives as a fair representation of what constitutes a true approach to solving the “Yarbus problem.” Namely, that the problem only applies to fixation and saccade data. The issue with this assertion, as pointed out by Haji-Abolhassani & Clark (2014), is that Borji & Itti did not use only fixation and saccade data/statistics, which is in conflict with what the reviewer later asserts is the reason why our study does not relate to the “Yarbus problem”. In the Borji & Itti (2014) study, series of fixations and saccades are analyzed in addition to spatial information which was later inferred from the eye movement data. It is unclear to us why this approach would be considered a more valid solution to the inverse Yarbus problem than our approach using minimally processed data.

Furthermore, there are other examples outlined in Table 1 our manuscript that use pupil size (MacInnes et al., 2018), eccentricity, and screen coverage (Krol & Krol, 2018), and more. These papers were not mentioned by the reviewer, so we are unable to discern whether they believe that these are also valid approaches to the inverse Yarbus problem. Given the lack of commentary with regard to these approaches, we have no reason to believe that they also do not qualify as a potential solution to the inverse Yarbus problem. This fits with our understanding that rather than applying only to eye movement data processed into a particular set of features (i.e., series fixations and saccades), the requirement for a valid approach or solution to this problem only needs to use eye movement data to infer the task-at-hand.

To mitigate future misunderstandings along these lines, we have clarified our interpretation of the inverse Yarbus problem where relevant in the manuscript.

1. Definition of “pattern”

Related to the previous comment, the reviewer chose to present a quote from Greene et al., and chose to make the assertion that the word “pattern” in the quote was referring specifically to the use of “series fixations and saccades, not raw data.” As we understand, from this quote specifically, pattern is referring to the earlier mention of “eye movements” in the Greene et al. quote, which actually does not refer to any type of data, but the movement of the eyes. To clarify this point, we present the quote initially brought up by the reviewer, including the sentence immediately following:

“Yarbus argued that changing the information that an observer is asked to obtain from an image drastically changes his pattern of eye movements. Moreover, the scan paths from this famous figure have been taken as evidence that eye movements can be windows into rather complex cognitive states of mind.”

As we can see, the phrase “pattern of eye movements” is most clearly in reference to a qualitative evaluation of the scan paths in the well known Yarbus figure named “Unknown Visitor,” which is presented in Greene et al. (2012) and many other articles.

Additionally, to further contextualize our response to the comments of Reviewer #1, we would like to include an additional quote from Greene et al. (2012) setting up the purpose and specific contribution of their paper:

“While these later studies have shown that an observer’s task can change certain individual features of eye movement patterns, they have not shown whether these differences can be used to identify the task of the observer. This has been an untested, but popular inference from the Yarbus finding as the visual differences between scan paths appeared so different.”

This further supports our assertion that Greene et al. were referring to the qualitative differences in the scanpaths for the different tasks. This quote also differentiates the “Yarbus problem” and “inverse Yarbus problem” by describing in the first sentence that task has been shown to influence eye movements, but that [at the time] there was no evidence that differences in eye movements between tasks could be used to identify the task-at-hand.

After reviewing the assertions of the reviewer, and the source material for these assertions, we have determined that our interpretation of the inverse Yarbus problem appears to differ from that of Reviewer #1. We believe that (a) there is an important distinction between the Yarbus problem and the inverse Yarbus problem, and (b) valid solutions are not required to only use eye movement data processed into particular forms. Our stance on these positions are supported within the relevant literature, and even the specific examples brought up by the reviewer. For this reason, we believe there is room in our manuscript for more clear communication of these ideas.

As stated in our other responses to Reviewer #1 comments, we have made an effort to avoid future misunderstandings by using more specific language where relevant in the manuscript.

1. Raw data

One more clarification that we felt was important to make is that we did not classify raw data. Rather, we used minimally processed data. This is an important distinction, namely because one of the important contributions of this manuscript was using a deep learning CNN to classify images of eye movement data. In order to develop these images, the eye movement data must go through some minimal amount of processing. We have added clarifying language to relevant portions of the manuscript in an attempt to mitigate any future misunderstandings in this regard.

(b) The reviewer provides two potential solutions that we address separately below:

1. Compare raw eye traces to the processed saccade and fixation data.

While this is a compelling approach, the solution is more complex than may be evident on the surface. Processing the data into fixation and saccade features/statistics would change the structure of the data, which would require an entirely new model. For this reason, a comparison of an aggregate approach and our minimally processed black box approach would not provide any insight as to whether the features emphasized in any of the deep learning models are the same or different.

Alternatively, a comparison of classification accuracies for our approach and another approach that processed the eye movement data into series fixation and saccades would provide a relative index of how well our approach answers the inverse Yarbus problem compared to other approaches documented in the literature. This is relevant to a specific request from Reviewer #2 (see Reviewer #2: #Other Points: 1). For this reason, we have obtained the code for the data processing and classification approach originally implemented by Coco and Keller (2014). It is worth noting that this method was the best performing solution to the inverse Yarbus problem, yielding upwards of 84% classification accuracy. Following Coco and Keller, we processed our data into seven of the features they used, and classified these features using the SVM model they presented. The results from this analysis approach were compared with the results from our minimally processed approach and discussed.

1. Rewrite the manuscript without the Yarbus problem framing.

The reviewer believes this would “make a good contribution.” Given our decision to compare our approach with the highest performing approach to the inverse Yarbus problem documented in the literature, we have chosen to continue to frame our results as a new approach to the inverse Yarbus problem. We feel that our paper is presented in a form that provides a relevant and beneficial contribution to the field (which appears to be in agreement with the comments made by Reviewer #2). As stated earlier, we believe that the original framing of our approach, specifically with regard to the Yarbus problem and the inverse Yarbus problem could benefit from clarification, which has been addressed throughout the manuscript.

Reviewer #2 (Comments for the Author (Required)):  
  
In the present manuscript, the authors apply deep learning to predict observer task from eye movement data. They collect multiple datasets where observers viewed images while they had to do one of multiple possible tasks. They test two different modeling approaches. The first one is a DNN operating directly on timelime data of x and y position and pupil size. The second approach first encodes the timeline data into images of gaze traces and then applies a DNN to these images. The authors find that both models can predict observer task with above chance accuracy, however, the image-based model has substantially lower performance than the timeline based model. In ablation analyses they find that x position is most important for task classification, while pupil size is least important.  
  
The effect of tasks on eye movement is very interesting, as can be seen from the long history of research on it. The paper is working on a very relevant question and is written in a very clear and understandable way, which I very much appreciate.  
  
In the following, I am going to list some conceptual and technical issues that I see. Once they are addressed, I think this work can be very relevant for the community.  
  
  
# Conceptual points  
  
1. First, I was a bit puzzled by the use of the terms "task", "cognitive process" and "mental state". In the paper, they seem to be used mainly interchangeably, but I think I would not agree with that. For me, intuitively, the task defines the goal of an observer, the cognitive process is the cognitive part of how the observer tries to reach this goal and the mental state is where in this process the observer is at a given time. Obviously, the task influences the cognitive process and the task will always be part of the mental state, but to me, cognitive processes and mental states carry much more information than the task. To me, all experiments and results in the paper seem to be about observer task, so I think conclusions about cognitive processing and mental states need more discussion.

We agree that the terms “task,” “cognitive process,” and “mental state” are all related, but do not carry the same meaning. To build off of what the reviewer has posited in their comment, eye movement features are instantiated by, and can be indicative of, the cognitive processes the cognitive processes instigating diagnostic patterns (i.e., features) in the eye movement data. These features are extracted and decoded by the deep learning CNN to determine the task-at-hand. This means that cognition is inferred from eye movements. Some cognitive processes can be associated with particular eye movement features, which can also be associated with particular tasks. Each task is assumed to be associated with a particular mental state.

The comments by Reviewer #2 reflect an issue that was largely touched on in our response to the comments from Reviewer #1. In particular, we agree with Reviewer #2 that this language can be further clarified in the manuscript. For this reason, we have gone through and updated the manuscript where necessary in order to clarify the distinction between the three terms.

2. The authors do not pass any stimulus information to their models. In the introduction, they argue that "such efforts to not fit the spirit of the inverse Yarbus problem, which is concerned with decoding high-level abstract mental operation that are not dependent on particular stimuli". I'm not sure whether I can agree with this. I would argue that the high-level abstract mental operation operates on the content of the viewed image: to guess how wealthy the people in an image are, I need to figure out where the people are in the image, I will likely inspect their clothing and other attributes and then process this data to come up with a guess of their wealth. Therefore, the eye movements are highly dependent on the stimulus itself. Without access to the stimulus, I don't see how one could expect to decode the mental operation that is going on. Of course there will be differences in the pure eye movements (as can be seen in the above-chance performance of the model), however they might be best explained in combination with the image viewed. For example, if the relevant areas of the image for the task at hand are more scattered over the image, longer saccades might be the result. Another example is given by the authors in the discussion, line 470: pupil size might be mainly affected by stimulus properties. Overall, I'm not sure whether one can expect to decode the "mental operation" or the "cognitive process" purely from the gaze data. However, this might boil down to my first point: Maybe I'm misunderstanding what the authors mean when they talk about mental operations etc.

We agree with the reviewer that there is likely some misunderstanding of the message we are trying to get across with regard to the spirit of the inverse Yarbus problem, and with regard to the use of stimulus information in the classification of eye movement data.

In reference to the quote presented by the reviewer, we were referring to previous research which has decoded eye movement data collected while participants were carrying out tasks such as reading vs. evaluating an image. In this case, reading requires mostly horizontal (left-to-right) eye movements following lines of text. For this reason, the scan paths of reading are very obviously qualitatively different from memorizing or evaluating an image. These low-level distinctions are so easily differentiated that the more subtle high-level distinctions between these tasks are ignored. In these cases, the tasks are tied directly to properties of the stimulus rather than properties of the cognitive processes differentiating the tasks-at-hand. This is what we feel doesn’t fit the spirit of the inverse Yarbus problem.

We do agree with the reviewer that stimulus information could improve the decodability of the processed images. This is actually mentioned in our Discussion section as a potential future direction to take with this line of research.

To more clearly communicate our intention to distinguish between studies that do and do not fit the spirit of the inverse Yarbus problem, we have further clarified portions of the manuscript relevant to this topic.

3. A major result of the paper is the comparison of the timeline model and the image model, where the authors find that the timeline model works much better for task classification. I'm not sure I fully understand what the authors hope to show with this. In theory, both models have access to exactly the same data (except for a few occlusions that should be inferable), so with sufficient computational capacity and training data, both models should reach exactly the same performance. Therefore, the differences in performance are based in different inductive biases of the different architectures and in training and overfitting problems. If I'm not mistaken, the image model has more than 100 times as many parameters as the timeline model and therefore can easily suffer more from overfitting. In addition, it is trained from scratch. I could imagine that adding a ImageNet pretrained backbone such as ResNet or DenseNet can improve performance (I'm aware that the input images are quite far from natural images, but even on out-of-domain data, deep models often still encode surprisingly useful features). However, besides all of those points, I can imagine that it is very hard to bring the image based model to the same performance as the timeline model. After all, the features that so far have been most successful in predicting task (see Table 1 of the manuscript) are much easier to compute from timeline data than they are to compute from the image data. Of course, if stimulus information should be included in the models, then the image-based model makes a lot of sense. Overall, I think the paper would profit if the authors state more clearly what they hope to learn from comparing these two models.

Deep learning CNNs are known for their proficiency in decoding image data. We were originally planning to process the data into images in an attempt to take advantage of the CNN’s ability to decode data in this form. We also classified the data in the timeline format because this is more typical for this field of research, and to provide a relative baseline for the image classification. Given the inability to truly compare the accuracy of our models to those of another study (a point that is covered in the manuscript, and in #Other points: 1), especially given the unique data features used in the current study, the comparison between image and timeline formats seemed like a relatively effective baseline.

We mention these points in the manuscript, but do not explicitly state the reason for the comparison of these model types. Clearly, the points made above have not been translated in the manuscript in a way that is clear to the reader. In an effort to clarify the reason why we chose to compare the image and timeline formats, we have added further description of the reason why we chose to process the data into image and timeline formats, and to then compare the classification accuracies of these two data types.

# Other points  
  
1. In the introduction, lines 150-163 the authors argue that due to different datasets used, it is hard to compare the different approaches used for classifying task in the past. I fully agree, however, I think this would be even more reason to check at least one or two of the better performing methods from Table 1 on the collected dataset. The blackbox DNN should outperform them since it has access to the full raw data (see also in the discussion, lines 421-423).

As stated in our responses above, we agree with the reviewer’s suggestion that it would be helpful to clarify the efficacy of our approach relative to others. At this point, it would be hard to say whether access to the full raw data would actually lead to better performance of our approach compared to another. This is because the other approaches lead to data that is processed into a different structure, thus requiring a different CNN model architecture. It could be that data processed in the format of the other approaches is more indicative of the task-at-hand and are more differentiable between tasks.

To address this issue, we obtained the code used by Coco and Keller (2014) to process and classify the data their eye movement data. Using the Exploratory dataset, we processed the data into seven features used by Coco and Keller (i.e., initiation time, number of fixations, mean entropy, mean saccade amplitude, mean fixation duration, percent area fixated, mean fixation saliency). These features were classified using all three model approaches used by Coco and Keller (SVM, LASSO, MM). As suggested by the reviewer, our approach outperformed Coco and Keller’s. The drastic drop in performance of the Coco and Keller approach between their study, and a re-analysis of our data using their approach is likely due to the explicit efforts made by Coco and Keller to differentiate the tasks in a way that would not be dependent on the analysis approach, which they demonstrated by using Greene et al.’s approach to successfully classify their data (Greene et al.’s approach failed on their own data). Our dataset contains tasks that are capable of being distinguished (as we, and others using the same tasks have demonstrated), but these tasks are not as easily distinguishable as the tasks used by Coco and Keller. For this reason, to successfully classify our dataset requires a sophisticated analysis approach. Although the Coco and Keller approach clearly demonstrated proficiency, outperforming the rest of the literature, the evidence provided here showing that our approach was able to outperform that of Coco and Keller provides further evidence (in addition to that already provided in the original manuscript) that our approach is state of the art. This process and the results have been described at length in the revised version of the manuscript.

2. In the ablation studies, the authors use datasets where one or more components have been removed (X, Y, P). From the manuscript, I'm not sure whether the models where retrained on the new data, or whether the already trained models where evaluated on the reduced data. For assessing the relevance of the different components, I would argue that the models should be retrained. Zeroing out some components introduces a substantial domain shift and model performance might just drop because of this.

In most cases, the models were retrained. The models were only not retrained in the supplementary analysis comparing the three tasks because independent subsets of the original results were calculated and compared. To clarify which models were re-trained, and which models were not re-trained, and the specific reasons why some models were not retrained, the manuscript was updated where appropriate.  
  
3. On a related note, if I see it correctly, for the generalization test from exploratory to confirmatory data, the models are completely retrained. Here, I think I would have gone for not retraining the models since the distribution shift is much smaller and it is actually interesting to see how well the confirmatory data can be predicted from the explanatory data (as opposed to other confirmatory data). But I don't request that the authors do that, I just wanted to point out that here, both approaches make sense. They are just answering different questions.

We can appreciate the reviewer’s comment on the decision to retrain the confirmatory model using the new dataset. We chose to retrain the confirmatory model because we felt like this was the best way to validate the efficacy of our overall approach, not just the model we used. While the model is certainly an important aspect of the study, we feel that the unique method of decoding the scan paths as images, and using the minimally processed datasets is the most interesting contribution that our manuscript brings to the field.  
  
4. Training/test split (line 270): If I understand this correctly, for each iteration, a new random training/test split was sampled. This seems a bit unusual to me, in deep learning usually either a fixed train/test split is used multiple times to assess the variance of the different random initializations of the model, or, sometimes, a full k-fold crossvalidation is used to make best use of all available data (but then error bars might be less relevant). Now the error bars will be partially due to the different initializations and partially due to the fact that different datasets were evaluated, that might be slighly different in their difficulty.  
  
What am I supposed to say here??

# Minor points  
  
\* please elaborate a bit on the relative computational capacities of the two models, i.e. the parameter counts. Was overfitting a problem for the image data model?

Details regarding the parameter counts have have been added to the model description in the manuscript.

\* I'm missing some information on the used learning rates. Was there no learning rate decay at all? I would expect the models to profit from using some sort of learning rate schedule (even when using ADAM).

Actually, there was a decay in learning rate as classification accuracy improved. This information has been added to the model description in the manuscript.

\* "To determine the relative value of the contribution from each component", [ANOVA was performed] (lines 296): This is really only a subtle point, but I would argue that ANOVA only determines whether there is any effect at all but not the relative value of each component. The relative value could be measured, e.g., by the differences in performance. Obviously, a statistical test is required to assess that these differences are meaningful, but ultimatively, in my opinion, it is more interesting to see how large the effect is.

The reviewer states here that the comparisons we present in the manuscript do not represent the relative value of the contribution of the components compared, but then states that the relative value could be measured by a difference in performance, which is something that can in fact be assessed using an ANOVA. This is the statistical analysis we use for our comparisons. We also present effect sizes with our comparisons as a way to show how large the effect is (i.e., the magnitude). For these reasons, we are not entirely certain what the reviewer is proposing with this comment, or if any changes to the manuscript are being suggested with this comment.

\* Figure 4: the dataset without pupil data seems to result in slightly larger performance than the full dataset. This is within the margin of error, but I wonder whether it indicates any overfitting problems (although it should not reduce the effective capacity of the model substantially). It might be interesting to discuss this result at least in a few words.

We agree that this is a worthy topic of discussion. We have explicitly added discussion of this issue in the manuscript.