Decoding binary decisions under differential target probabilities from pupil dilation: A random forest approach

Christoph Strauch Experimental Psychology, Helmholtz Institute, Utrecht University, the Netherlands

Teresa Hirzle Institute of Media Informatics, Ulm University, Germany

Stefan Van der Stigchel Experimental Psychology, Helmholtz Institute, Utrecht University, the Netherlands

Andreas Bulling Institute for Visualisation and Interactive Systems, University of Stuttgart, Germany

Although our pupils slightly dilate when we look at an intended target, they do not when we look at irrelevant distractors. This finding suggests that it may be possible to decode the intention of an observer, understood as the outcome of implicit covert binary decisions, from the pupillary dynamics over time. However, few previous works have investigated the feasibility of this approach and the few that did, did not control for possible confounds such as motor-execution, changes in brightness, or target and distractor probability. We report on our efforts to decode intentions from pupil dilation obtained under strict experimental control on a single trial basis using a machine learning approach. The basis for our analyses are data of 69 participants who looked at letters that needed to be selected with stimulus probabilities that varied systematically in a blockwise manner ($n = 19{,}417$ trials). We confirm earlier findings that pupil dilation is indicative of intentions and show that these can be decoded with a classification performance of up to 76% area under the curve for receiver operating characteristic curves if targets are rarer than distractors. To better understand which characteristics of the pupillary signal are most informative, we finally compare relative feature importances. The first derivative of pupil size changes was found to be most relevant, allowing us to decode intention within only about 800 ms of trial onset. Taken together, our results provide credible insights into the potential of decoding intentions from pupil dilation and may soon form the basis for new applications in visual search, gaze-based interaction, or human–robot interaction.

Introduction

If the eyes are a window to the soul, or – more specific – reflect the intention of an observer, how far open is this window and which features let us see through it? Because the pupil dilates more for targets than distractors (de Gee, Knapen, & Donner, 2014; Einhäuser, Koch, & Carter, 2010; Privitera, Renninger, Carney, Klein, & Aguilar, 2010; Strauch, Koniakowsky, & Huckauf, 2020), pupil dilation is one of the most promising features to indicate whether a foveated stimulus is an intended target. Therefore, analyzing the pupil, possibly even in real time, could allow for decoding such intentions, be it by human or machine. But how well may intention, or more precisely the outcome of a binary decision, be decoded by observing pupil dilation? Although findings in pupillometry are commonly reported as average or maximum changes compared to a baseline, pupil characteristics that are not easily visible from graphs to the naked eye are likely to carry additional information. Machine learning methods can reveal whether intention can indeed be decoded from pupil dilation and, as such, help us to better understand how much information our pupils reveal to the outside world. Furthermore, the analysis and comparison of a set of potentially informative features of the pupil size signal may allow us to understand which signal components carry most information and can hence be considered worthwhile to be investigated with other types of analysis. Although the only two existing studies in this direction with a dedicated focus on pupil size (Bednarik, Vrzakova, & Hradis, 2012; Medathati, Desai, & Hillis, 2020) have

Citation: Strauch, C., Hirzle, T., Van der Stigchel, S., & Bulling, A. (2021). Decoding binary decisions under differential target probabilities from pupil dilation: A random forest approach. *Journal of Vision*, *21*(7):6, 1–13, https://doi.org/10.1167/jov.21.7.6.

https://doi.org/10.1167/jov.21.7.6 Received February 15, 2021; published July 14, 2021 ISSN 1534-7362

set the path, a number of limitations leave the initial questions largely unanswered. The present investigation seeks to address these limitations and provide an exhaustive answer into the principal possibility of decoding intention from pupil dilation.

Theoretical background

Previous studies have found that the pupil cannot only identify when we decide (Einhäuser, Stout, Koch, & Carter, 2008), but also which decision was reached (Hakerem & Sutton, 1966). When aligning pupillary responses to fixations during a visual search, larger dilations were found on targets compared with elsewhere (Klingner, 2010; Martin, Whittaker, & Johnston, 2020). Furthermore, pupils dilated stronger when targets were presented with fixed gaze in rapid serial visual presentation protocols (Privitera et al., 2010). Similarly, in binary decision tasks using signal detection paradigms, pupils were found to dilate more whenever participants *thought* that a signal was present compared with just noise instead of the physical presence of a signal (de Gee et al., 2014). In another task, participants were sequentially presented with numbers and reported which one they mentally selected afterward. The results showed that these numbers were associated with a slightly larger pupil dilation than rejected numbers (Einhäuser et al., 2010). Two more applied scenarios, one including the selection of letters on an on-screen keyboard, found pupils to dilate relative to a directly foregoing local baseline when to-be-typed letters were fixated (Strauch, Ehlers, & Huckauf, 2017, Strauch, Greiter, & Huckauf, 2017). When participants had to look at letters either matching or mismatching a target letter (with the effects of eventual key presses being controlled), pupils dilated stronger for target than for distractor stimuli (Strauch, Greiter, & Huckauf, 2018).

One of the most important determinants for the difference found between targets and distractors, be it in visual search, decision-making tasks or applied scenarios, is the relative proportion of targets to distractors (Strauch et al., 2020). Specifically, the rarer a stimulus the stronger the pupillary response relative to more frequent stimuli – a phenomenon commonly referred to as the oddball effect (Murphy, O'connell, O'sullivan, Robertson, & Balsters, 2014). In a recent study, the effects of binary decision-making were investigated together with effects of stimulus probability (Strauch et al., 2020). Although targets always elicited a stronger pupillary response than distractors, the magnitude of the difference between targets and distractors was determined by an effect of stimulus probability. Targets elicited a substantially stronger response when they were rare relatively to being equiprobable or frequent. Distractors, however, were always associated with the same response, suggesting that the effect observed may be attributed to differential activation elicited by targets depending on probability. Hereby, probability might be affecting the physiological activation by modulating a target's relevance, that is, effects should be strongest when targets are highly relevant to participants (Strauch et al., 2020). To allow conclusions for a broad range of setups, stimulus probabilities must be considered in systematic investigations of how well the pupil may predict intention. For example, if predicting intention in a binary decision task, targets and distractors might be equally probable, whereas in a visual search task, targets might be rare by default. Summing up, there is a growing body of relatively controlled experiments demonstrating a stronger pupil dilation in response to target compared with distractor stimuli; however, all these reports are based on the aggregation of multiple trials of multiple participants. This factor raises the question whether the outcome of individual decisions could be decoded from pupil dilation.

Addressing the question whether pupil dilation may decode decisions on a single trial level, a few investigations using support vector machine classifiers (SVMs) have been put forward over the past years (Bednarik et al., 2012; Jangraw, Wang, Lance, Chang, & Sajda, 2014; Medathati et al., 2020). Although several articles have found random forest classifiers to be particularly useful for predictions also based on pupil dilation (Kootstra et al., 2020; Pasquali et al., 2020), no prior work has focused on predicting intentions only using pupil size. Jangraw et al. (2014) used electroencephalography and ocular parameters to infer objects that were of subjective interest to users in a virtual reality environment and found both to be predictive. During informational intent, defined as search for an object under the assumption that it exists, the pupil was found to be larger than during scanning of a scene. Together with other gaze features, such as fixation durations, an accuracy of over 85% could be reached using an SVM (Jang, Mallipeddi, Lee, Kwak, & Lee, 2014). Bednarik et al. (2012) explored the potential of pupil dilation and other gaze characteristics as a means to infer the intention to select one out of the possible puzzle tiles in an eight-tile slide puzzle. A classification performance of about 60% was reached using models with SVMs when considering pupil dilation only; by including further gaze characteristics, classification performance reached up to 80% (Bednarik et al., 2012). Using pupil sizes obtained from a variety of tasks, a classification of about 60% was reported for changes in cognitive state, which in most (but not all) cases likely reflected a binary decision on a target's presence in visual search. Hereby, models using a SVM for rolling windows of 1 to 2 seconds achieved the best results when data were locally z-standardized

(Medathati et al., 2020). However, the considerable differences between the tasks leave unanswered what classification could be reached for binary decisions in particular. Deriving with valid conclusions from pupil dilation usually requires highly controlled experiments, as also motor execution, such as a key press (Richer & Beatty, 1985), changes in gaze position, if not corrected for Hayes and Petrov (2016), and brightness affect pupil dilation. In the foregoing investigations most closely related to recognizing user intention, understood as the outcome of a binary decision on a stimulus' relevance from pupillometric data, changes in gaze position were not controlled for (Bednarik et al., 2012; Medathati et al., 2020). More important, however, key presses were necessary to inform about the intent to select. Pressing a key alone leads to strong pupil dilations (see Figure 3 for the effect of a key press in this investigation), usually exceeding effect sizes of binary decisions (Richer & Beatty, 1985; Strauch, Greiter, & Huckauf, 2018). Thus, the pupil does not only pick up on a change in activation elicited by the intention (de Gee et al., 2014), but also on the substantially larger change elicited by motor execution. In other words, it cannot be fully excluded that the observed effects were not exclusively driven by the decision regarding a stimulus' relevance, but the motor execution that was associated with the very same decision in the analyzed tasks. Hence, previous studies using SVM algorithms (Bednarik et al., 2012; Medathati et al., 2020) remain inconclusive as to whether *intention alone* can be decoded from the pupillary signal. Machine learning features solely derived from pupil size can further foster the understanding which signal characteristics are most informative and thus most promising at the center of pupillometric methods and investigations.

Previous work thus raises the following, not exhaustively answered research questions:

- How well may intention be decoded from pupil size?
- How is the classification accuracy affected by stimulus probability?
- Which components of the pupillary signal are most informative about intention?

To answer these questions, we reanalyzed pupillary data that was obtained for the investigation into binary decision-making and stimulus probability described elsewhere in this article (Strauch et al., 2020). The original investigation focused on the interaction between decision-making and stimulus probability and the sequentially of respective effects (Strauch et al., 2020) in a strictly controlled setting using averaged pupil sizes for inferences. The acquired data are, however, also well-suited for investigating the feasibility of decoding intention from pupil size on a single trial basis. Furthermore, the data allow to compare differential features derived from pupil size regarding their information content in an optimal manner, due to the strict experimental control of effects of gaze position, brightness, and motor execution across conditions. We use a random forest approach to classify pupil size as intent or non-intent related using a dataset consisting of more than 19,000 trials, which was gathered from 69 participants.

Methods

Two experiments were conducted that were similar unless stated otherwise. All data may be retrieved together with the analysis scripts and Supplementary Material via the open science framework (https://osf.io/xezf4/).

Participants

A total of 69 participants took part in the experiments (experiment 1: $n = 44$; experiment 2: $n = 25$). All participants reported normal or corrected to normal vision with contact lenses and gave written informed consent before the study. The study was approved upfront by the local ethics board, adhering to the Declaration of Helsinki.

Apparatus

Pupil dilation was obtained using a SMI Hi-Speed 1250 Eye tracker (SensoMotoricInstruments GmbH), which features a chin rest; pupil data were obtained at 50 Hz. A 27-inch screen with a resolution of 1920 × 1080 pixels and a 144 Hz refresh rate was positioned at 60 cm from eye position. Brightness was kept constant at 60 lx at eye position. A standard keyboard was used to register key presses.

Design and procedure

The factors decision (target/distractor) and key press (with/without) were varied in all experimental blocks. Stimulus probability was varied block-wise with 25% targets, 50% targets, and 75% targets in Experiment 1 plus 25% targets and 75% targets in Experiment 2. Participants were instructed by a written instruction and the experimenter that they had to look at letters and check whether these match a given target letter in the experiments after written informed consent was obtained and the eye-tracker was calibrated.

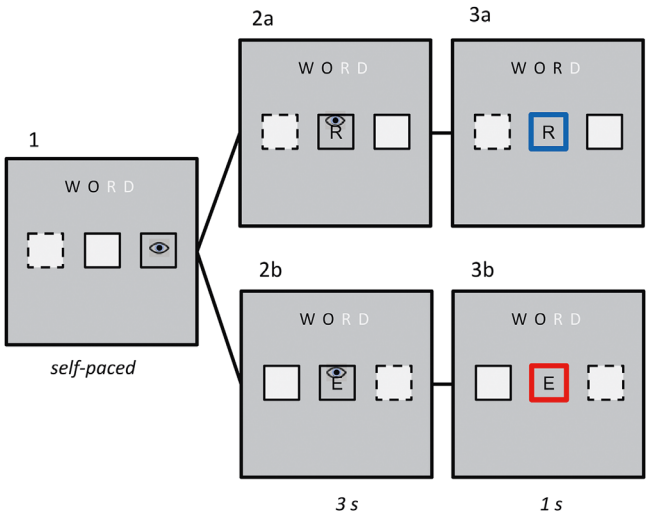

Figure 1. A schematic flow chart of possible trial configurations. A neutral four letter target word was always presented in the upper screen center. Participants could start a trial by first looking at a box that was randomly presented either right or left of a central box; successful activation of a box was associated with its color turning to the background gray. Once the central box was activated and gaze kept inside, a letter was presented centrally that either matched (target) or mismatched (distractor) the leftmost gray letter above for 3 seconds. Upon completion of a trial, edge lines turned red for distractors and blue for targets for 1 second. For the latter, the respective letter in the target word above turned black to indicate the next target letter. Participants could start a new self-paced trial.

Task

A schematic depiction of the task and possible trial configurations is provided in Figure 1. A gray screen with a light gray box in the center with black edge lines of 4.5 visual degree edge length was presented. A box of similar properties was presented at 4.5-degree visual angle randomly either right or left to the central box. In the upper screen center, a neutral word with a length of four letters was given. Trials were started by first looking at the outer box and then saccade to the central box (Figure 1). Boxes changed their color to the gray of the background, except for the black edge lines, once gaze position was registered inside to provide feedback. Once participants entered the central box by gaze, a letter was presented inside (Figure 12). This letter could either match (target; Figure 12a) or mismatch (distractor; Figure 12b) the leftmost gray letter from the word presented above. In both cases, participants were instructed to keep their gaze position inside the central box for 3 seconds to absolve trials correctly. Successful trial completion was fed back with blue edge lines (Figure 13a) of the central box for targets and red edge lines of the central box for distractors for 1 second (Figure 13b). For targets, the respective letter in the word above turned black and hereby indicated the next to be selected letter (Figure 13a). Participants could rest their eyes upon trial completion and start a new trial self-paced. Once all letters of the four letter word were "typed," the next four-letter word was presented. Participants were instructed that the experiment would end once all words were typed. The proportion between target and distractor letters was varied in blocks. Independent from the other manipulations, at a chance of 50%, a sine-wave tone (200 ms, beginning with trial onset) prompted participants to indicate via key press whether a presented letter was a target (right arrow key) or a distractor (left arrow key) during the trial. When participants left the central box early or blinked for longer than 200 ms, the trial aborted. An error was recorded whenever a key was pressed without the sound prompt, no key was pressed after the prompt, or an incorrect key was recorded. Errors required to retype the last two letters of the above presented word and thus prolonged the experiment. Given the relatively drastic consequence of errors, no competing stimuli, and that gaze needed to be kept in the central box, we assumed participants to pay attention also when no key press was required.

Results

Data preprocessing

Blinks were interpolated using the same algorithm as in Strauch et al. (2020), because only effects of pupil dilation (and not blinks) were in the center of the current investigation. There is no gold standard for preprocessing pupillary data; however, descriptive curves indicate the appropriateness of the chosen methods in that artifacts are very rare (see Figure 2). Pupil size changes throughout trials were z-standardized within participants to facilitate cross-sample comparisons. Figures depict pupil size changes in mm relative to the local baseline as a more intuitive measure.

Descriptive results

Weighted and averaged results from Experiment 1 and 2 of Strauch et al. (2020) are given in Figure 2A and Figure 3. As expected, the key press necessary in random 50% of trials had a descriptively stronger impact on pupil dilation (Figure 3) than the binary decision target/distractor (Figure 2A). In all conditions, pupils dilated stronger for targets than for distractors; this difference was strongly modulated by stimulus probability, that is, for rare targets, pupils dilated stronger than during equiprobability than for frequent

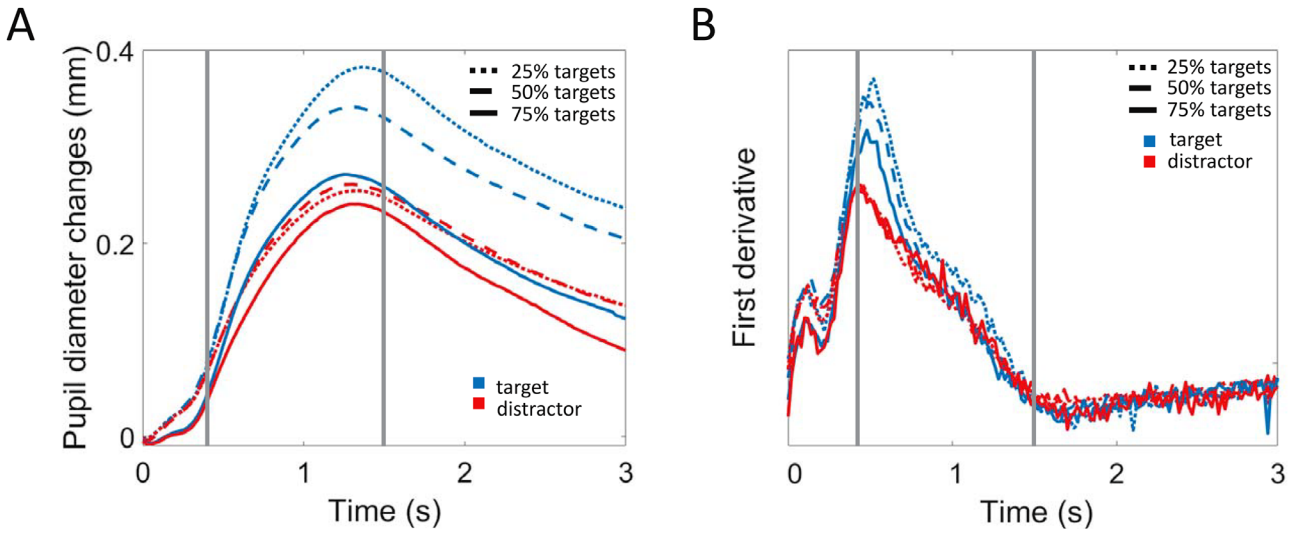

Figure 2. (A) Pupil size changes during trials relative to a foregoing local baseline of 20 ms, directly preceding trials. Targets (blue) dilated more than distractors (red). The difference between target and distractor was affected strongly by stimulus probability, with targets dilating strongest when they were rare (dotted), followed by equiprobable targets and frequent targets. Distractors remained mostly unaffected by stimulus probability. (B) First derivative/slope of the pupillary changes depicted in A. Note that the first derivative reveals the difference between target and distractor substantially earlier than the average. Gray vertical lines denote the beginning and end of the interval that was used for feature extraction.

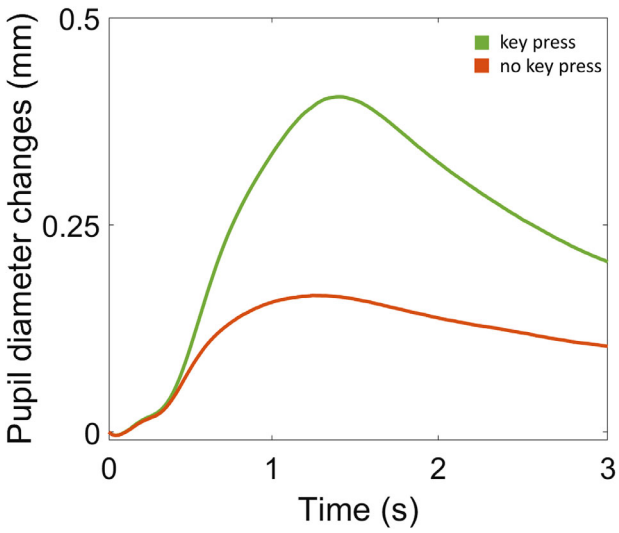

Figure 3. Pupil dilation relative to the foregoing local baseline for all trials with a key press (green) and all trials without key press (orange).

targets (Figure 2A). The first derivative of the average pupil courses reveals respective effects earlier than the average pupil size (see Figure 2B), that is, the effects are descriptively most pronounced already at 560 ms into trials. The first response visible right at the start of a trial until about 250 ms hereby likely reflects the orienting response and the processing of newly incoming information. The second response, roughly between 300 ms and 800 ms, shows highly similar responses to distractor letters irrespective of stimulus probability; rare targets led to faster changes in pupil size and pupils dilate longer compared with more frequent targets.

Model implementation

To systematically investigate whether differences in pupil dilation can be classified as intent or non-intent-related signals, we used a binary classification approach.

Feature extraction

As input for a binary classifier, a feature vector is extracted from the pre-processed raw signal, that is, the baseline corrected z-standardized pupil sizes with interpolated blinks. The feature vector consists of a number of features that represent informative properties of the signal. Identifying informative features is a challenging problem and automating this task has only recently become feasible with methods that learn rich representations directly from raw data (deep learning). Learning such representations, however, requires large amounts of training data, which are not available here. We therefore opted to follow the traditional approach of designing informative features by hand, guided by theoretical considerations.

To find suitable and relevant features, we first looked for parts in the signal that may hold information about the underlying cognitive processes. Within the 1-second sequence of the pupil signal, starting at 0.4 seconds,

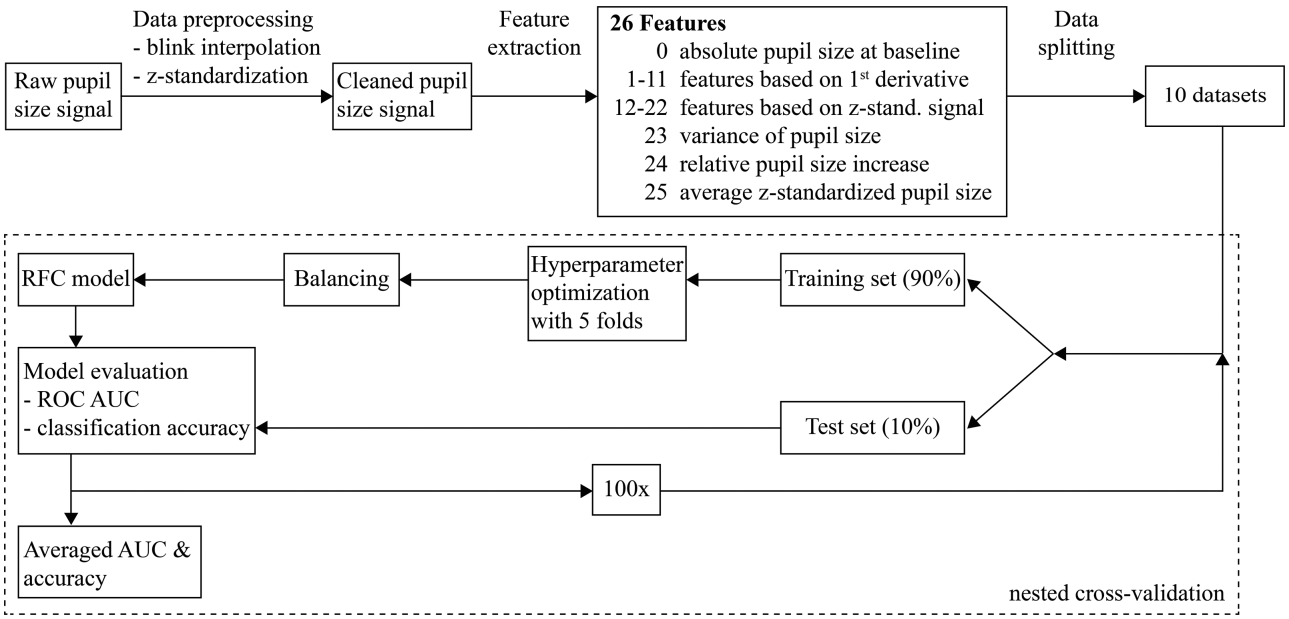

Figure 4. A representation of the signal processing and model implementation steps from the raw pupil size signal to the evaluation of the random forest classifier models. RFC, random forest classifier; ROC AUC, receiver operating characteristic area under the curve.

differences between conditions were descriptively largest (see Figure 2B) and therefore likely to carry the most relevant information for classification: In Strauch et al. (2020), the effects of decision-making were also statistically most pronounced between 0.4 seconds and 1.5 seconds after stimulus onset. Functional analyses, as performed in Strauch et al. (2020), can provide an understanding of how factors relate in their effect on pupil size and preserve temporal resolution while preventing the cherry picking of intervals for discretized analyses (Jackson & Sirois, 2009) and have proved their usefulness with pupillometric data before (Einhäuser et al., 2008; Jackson & Sirois, 2009; Strauch, Greiter, & Huckauf, 2018; Wetzel, Einhäuser, & Widmann, 2020). In contrast with machine learning classifiers, however, they cannot classify data into groups equally well.

Therefore, we focused on the initial increase in pupil size changes, but not on the complete decrease of the signal, to extract features. The first 0.4 seconds of the pupil signal were excluded owing the latency in the signal after stimulus presentation, derived from the descriptive data. We included 0.5 seconds after the average peak of pupil size in the signal. We considered the last 1.5 seconds (second 1.5–3.0) of the signal as irrelevant for the predictive model, given that differences between conditions remain descriptively constant (see Figure 2A).

Based on these findings, the following features were extracted for the interval from 0.4 seconds to 1.5 seconds of the z-standardized pupil size signal. First, we included the average pupil size throughout the interval as the standard measure of pupil dilation. We further calculated the maximum and minimum of pupil dilation within the defined interval, and included the difference as an indicator of pupil size changes during the trial as indicator of amplitude, another standard measure. As an indicator of pupil stability, we included the variance of the signal within the interval. Although the z-standardization helps to generalize findings and relative changes are usually in the center of investigations, information on absolute diameter is lost. Given that absolute pupil size has been associated with differential visual task performance repeatedly (Eberhardt, Strauch, Hartmann, & Huckauf, submitted; van Kempen et al., 2019), but z-standardization removes this information, we further included absolute pupil size in millimeters at baseline. To include features that preserve the temporal information, we sampled the z-standardized pupil size signal using a frame rate of 10 Hz and included the respective frames as features resulting in a number of 11 features. Last, as a marker for the velocity of pupil size changes, we included 11 features based on the first derivative, as these might show changes earlier than relative changes.

To obtain these features the first derivative of the pupil size signal was sampled at a frame rate of 10 Hz, similar to the previous time-related features. In total, we calculated 26 features, which were integrated into a feature vector that served as input for the prediction model.

	Experiment 1		Experiment 2	
Prop. target: Distractor	Key press	No key press	Key press	No key press
25:75	3,736	3,863	2082	2074
50:50	1,877	1,938	—	—
75:25	1,162	1,185	787	713

Table 1. Number of samples for each classifier. Notes: For Experiment 1, 39 participants were assigned to the training set and 5 participants were assigned to the test set. For Experiment 2 in the 25:75 group, 22 participants were assigned to the training set and 3 participants were assigned to the test set. For Experiment 2 in the 75:25 group, 21 participants were assigned to the training set and 3 participants were assigned to the test set. Which participants were part of either training or test set was randomly determined for each of the 100 repetitions of each model.

Data splitting

Because we used data from two experiments with different numbers of conditions and participants, we chose to build separate models for each experiment. Furthermore, because one goal of this investigation was to evaluate to what extent target probability has an influence on classification performance, we built separate models for each group of target probability. Last, we expected that a key press could have an additional effect on classification performance. Thus, ten models were built in total by splitting the total dataset by experiment (E1/E2), key press (with/without), and stimulus probability (25%/50%/75% targets for Experiment 1 and 25%/75% targets for Experiment 2). That is, in total, 10 individual binary classification problems were investigated (see Table 1 for the specific dataset splits).

To build a general model that is able to correctly decode binary decisions from pupil size signals, it is important to test the classifier on unknown data. Therefore, we used a nested cross-validation with 100 folds in the outer loop and five folds in the inner loop for each model. Therefore, for each model and each outer loop iteration, the total dataset was split into a training (90%) and a test set (10%). As such, it is ensured that the models were tested on data that they have not previously seen during the training phase. At this, the data was split by participants, that is, 90% of the participants were assigned to the training set and 10% to the test set. Training sets for Experiment 1 contained all data from 39 randomly selected participants. Training sets for Experiment 2 contained all data from 22 randomly selected participants for the 25% targets group and 21 participants for the 75% targets group due to missing data from one participant (see Table 1 for specific sample sizes for each dataset).

Classifier

To classify a feature vector as an intent- or non-intent-related pupil size signal we opted for the state-of-the-art random forest classifier for binary classification that has been used in prior work with eye tracking data (Kootstra et al., 2020; Pasquali et al., 2020). The implementation was based on the RandomForestClassifier of the sklearn.ensemble package of the Scikit-learn Python library (Pedregosa et al., 2011).

Optimization

The performance of a random forest classifier highly depends on a variety of hyperparameters that establish the model architecture, such as the number of trees in the forest or the maximum depth of the trees. The choice of hyperparameters is highly model dependent. Therefore, it is important to optimize the hyperparameters on the training set for each model before evaluating its performance on the test dataset. A grid search was performed for hyperparameter optimization for each of the outer loop folds of the ten models. For the hyperparameter optimization we used a five-fold cross-validation on the training set. A grid search is a comprehensive search over the specified values for the defined hyperparameters. Six hyperparameters were evaluated by the search: the number of estimators (400/800/1,000), the maximal depth of the tree (10/20/40/none), the minimum number of samples required to be at a leaf node (1/2/4), the minimum number of samples required to split an internal node (2/5/10), the function that is used to evaluate the quality of the split (gini impurity/entropy), and whether bootstrap samples are used when building the trees (y/n).

Balancing

A crucial factor for the prediction quality is the unbalanced distribution of samples in each class. This is especially important in our case, as we have an imbalance of classes of 25%/75% in eight of the cases. Therefore, class weights were adapted according to the prevalence of samples in each class (prevalence was given in percent).

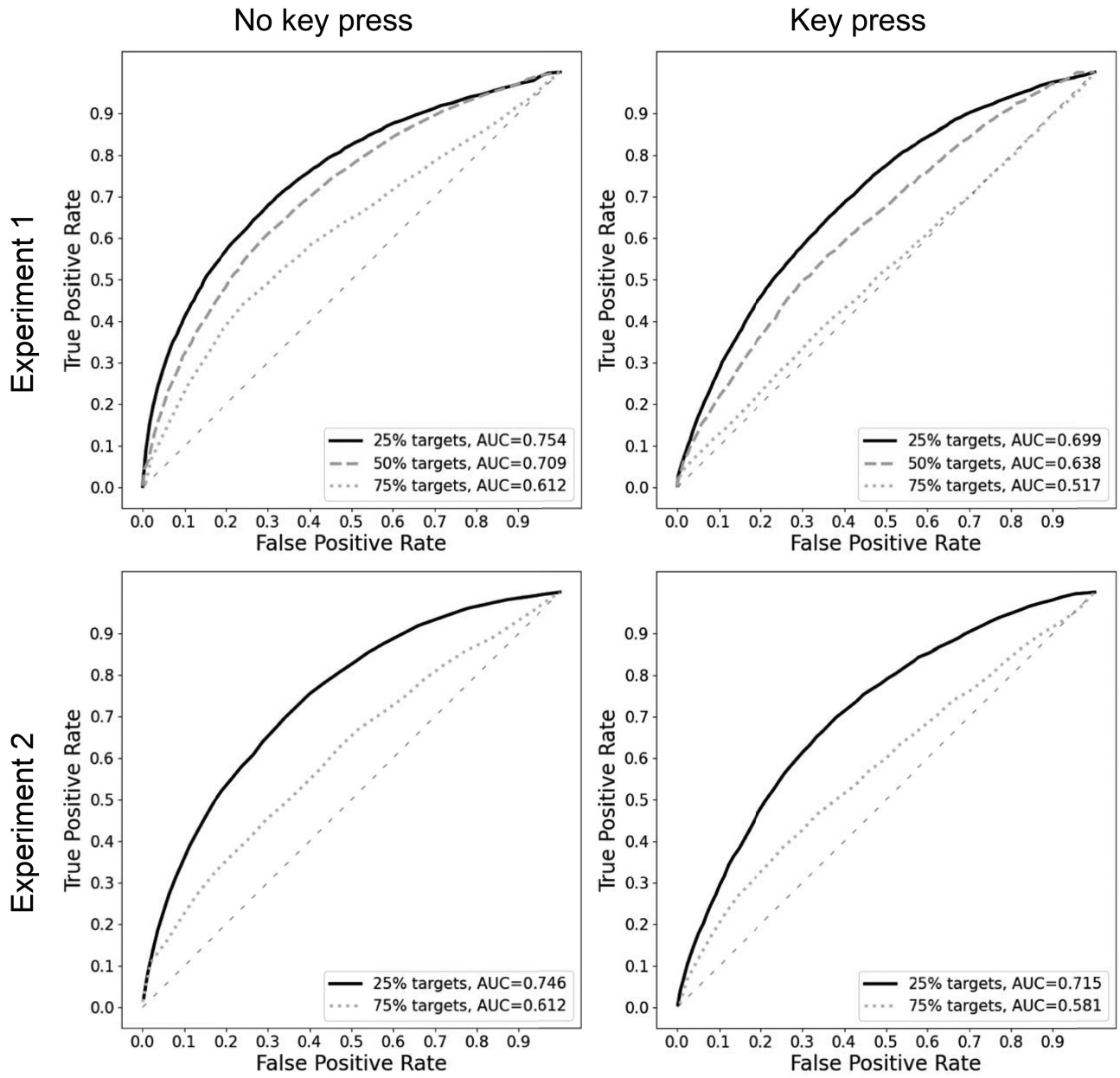

Figure 5. Receiver operating characteristic curves, separately for Experiment 1 (upper row) and Experiment 2 (lower row). Receiver operating characteristic curves for trials with key press are displayed on the left, without key press on the right. Solid lines for blocks with 25% targets, dashed lines for blocks with 50% targets, and dotted lines for blocks with 75% targets.

Model evaluation

Last, the models were evaluated based on the previously defined test sets. That is, the model was trained on each 90% of the data and tested on the independent test set, consisting of 10% of the data. To evaluate the performance of the models two types of evaluation metrics were applied. First, the classification accuracy score was calculated, a measure that indicates the mean accuracy of a classifier, defined as the percentage of correct predictions. Secondly, the receiver operating characteristic area under the curve (AUC) was calculated. The receiver operating characteristic curve is generated by plotting the true-positive rate on the y-axis and the false-positive rate on the x-axis.

Classification Results

The receiver operating characteristic curves of the classification results are shown in Figure 5. The accuracy and AUC for receiver operating characteristic curves values for all groups are given in Table 2. The best results were reached in the conditions with 25% probability for a stimulus to be a target, followed by the

	Experiment 1		Experiment 2	
Prop. target: Distractor	Key press	No key press	Key press	No key press
25:75	0.70 (0.14; 0.77)	0.76 (0.16; 0.80)	0.72 (0.06; 0.78)	0.75 (0.07; 0.77)
50:50	0.64 (0.04; 0.64)	0.71 (0.05; 0.65)	—	—
75:25	0.52 (0.05; 0.70)	0.61 (0.05; 0.71)	0.58 (0.07; 0.66)	0.61 (0.08; 0.70)

Table 2. Area under the curve for receiver operating characteristic curves separately for Experiment 1, Experiment 2, and key press/no key press conditions. Results are given separately for differential target:distractor probabilities. Values in brackets are standard deviation across runs and average accuracy.

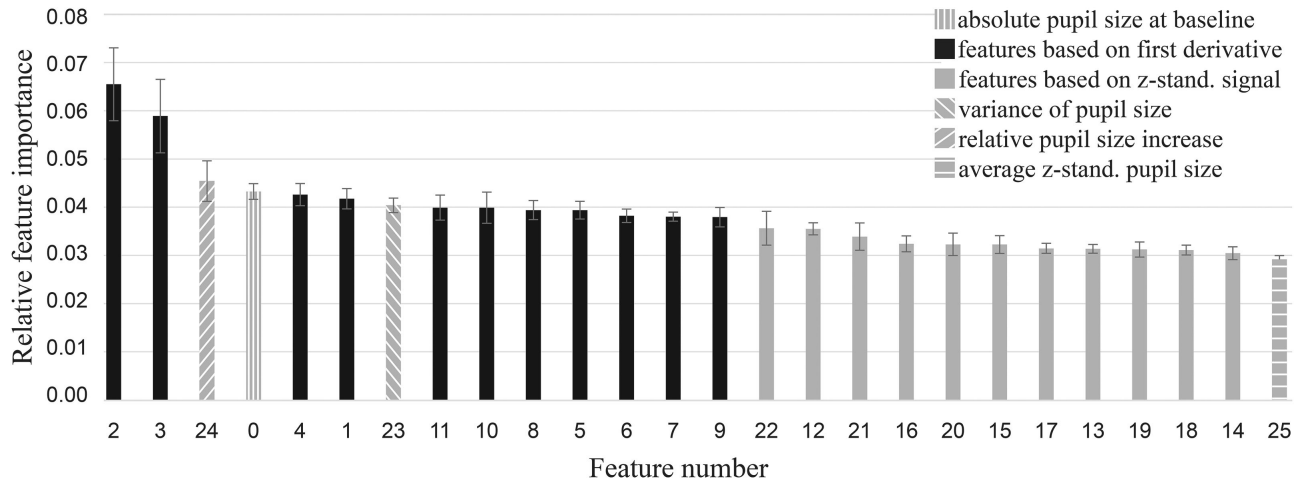

Figure 6. The relative importance of each feature averaged over all 10 classifiers (total features: 26). Feature 0 is the absolute pupil size during the first 20 ms of trials, features 1 to 11 reflect the sampled first derivative of the pupil size signal. Features 12 to 22 reflect the sampled z-standardized pupil size signal. Both were obtained by sampling the values with 10 Hz. Feature 23 is the variance of the z-standardized pupil values of the sequence 0.4 to 1.5 seconds. Feature 24 is the total pupil size increase (max-min) of the pupil size signal of sequence 0.4 to 1.5 seconds, and Feature 25 is the mean of the z-standardized pupil size signal in the respective sequence. Whiskers represent average standard deviations.

conditions with 50%. The worst area under the curve for receiver operating characteristic curves results were obtained in the conditions with 75% target probability.

Feature ranking

Figure 6 visualizes the averaged relative importance of features over all conditions. Most information was conveyed already by the first derivative of pupil dilation and absolute pupil size at baseline; further, the relative pupil size increase was found to be informative. In additional analyses employing just five features (Features 2, 3, 0, 4, 1), reflecting the first derivative of pupil dilation between 400 ms and 800 ms and absolute pupil size, the AUC was found to be only slightly worse than for the full feature set. Analysis demonstrates an AUC of 0.67 for the 25% target condition, an AUC of 0.65 for the equiprobable condition, and an AUC of 0.54 for the 75% target condition on average. Relative to the 25 features, this is 0.08 worse for the 25% targets condition and almost equal for equiprobability (0.03 worse) and slightly worse for the 75% targets condition (0.04 worse). See the Supplementary File via https://osf.io/xezf4/ for full results of this additional analysis.

Individualized models

These classifiers were trained and tested using data gathered from participants as part of either training or test set. This approach does not answer, however, whether classifiers can be tailored to persons, that is, whether idiosyncratic patterns can be learned that help decoding decisions. To address this question, we trained and tested classifiers separately within each individual, that is, a fraction of trials was used as training set with the rest of the set used as test set. Compared with the classifier trained and tested between participants, results were similar. Specific results of these analyses for all participants can be found in the Supplementary File and via https://osf.io/xezf4/.

Discussion

To investigate how well the outcomes of binary decisions may be decoded from pupil dilations that is, whether a processed stimulus is a target or a distractor, and to investigate which features covered in the pupillary signal are most informative for these classifications, we used 19,417 trials from 69 participants to train and test a random forest classifier. We strictly controlled for gaze position, brightness, and the effects of motor execution.

These results indicate that pupil dilation may indeed serve as a useful predictor for revealing intention in binary decisions with an AUC of up to 76% against a 50% chance level using a random forest classification approach. The AUC for classification was generally the better the rarer a stimulus . The difference between target and distractor became evident earlier from the first derivative than from z-standardized pupil sizes: The descriptive difference between target and distractor trials was maximal at 560 ms after trial onset and thus 580 ms earlier than for the average pupil size relative to baseline. Hereby, signal courses were almost perfectly aligned for distractors whereas they differentiated for targets. Thus, as indicated by Strauch et al. (2020) for average pupil dilation, stimulus probability seems to only play a role for processed targets, not all stimuli per se also when it comes to the acceleration of pupillary expansion. Note that the peak acceleration seems to be slightly lagged for targets relative to distractors, by descriptive tendency the more so, the rarer a target.

A comparison of predictors indicates that the early data points of the first derivative, the relative pupil size increase (i.e., the amplitude from baseline), and the absolute pupil size at baseline prove to be more useful than the average pupil size relative to baseline for the whole trial, and the overall pupil size relative to baseline where effects of decision-making were most pronounced, as well as the variance in the pupillary signal. Of course, it is also possible that these features just do not yield additional information, the first derivative, and absolute pupil size at baseline, however, provide it substantially faster. This finding is interesting to note for pupillometry, as most analyses focus on measures such as average pupil dilation over time (Ehlers & Meinecke, 2020; Naber & Murphy, 2020), maximum pupil dilation, or latency to peak pupil dilation (Koelewijn, Zekveld, Festen, & Kramer, 2012). Particularly short-lived cognitive processes or processes occurring in a relatively fast temporal succession could thus be addressed using the first derivative as an easy to implement and to analyze alternative to deconvolution techniques (Wierda, van Rijn, Taatgen, & Martens, 2012). This finding further provides the advantage that no baseline is required. Hence, the first derivative, could provide valuable information in combination with other analysis techniques, such as the aforementioned functional data analysis approaches (Jackson & Sirois, 2009; Strauch et al., 2020; Wetzel et al., 2020). Absolute pupil size in turn, could be added as a predictor to these functionally applied models.

For applications making use of real-time pupil-based judgments, this implies that fast information retrieval is possible, particularly when using the first derivative in combination with absolute pupil size at baseline, that is with up to 76%, within about 1.5 seconds. Interestingly, all most informative features but the relative increase in pupil size were available already as early as 800 ms, suggesting that the most information is already conveyed in this early pupil size change, suggesting the feasibility of even faster intention retrieval. Classifications were almost as reliable already using just these early features (see the Supplementary File for respective classification results). When targets were rare, participants were fastest in pressing a key, but this still took more than 950 ms on average (Strauch et al., 2020), suggesting that the overt communication of a decision is substantially slower than the here presented covert pupil-based decoding. For research, this offers the possibility to infer intention substantially faster using pupillometry than with overt manual responses at the cost of reliability.

Foregoing investigations have demonstrated that factors such as changes in brightness or motor execution primarily add to pupil dilation in a linear fashion (Pfleging, Fekety, Schmidt, & Kun, 2016; Strauch, Greiter, & Huckauf, 2018; Van der Stoep, Van der Smagt, Notaro, Spock, & Naber, 2021), which is why our results should also be of use when these factors are varying. For conditions including a key press, however, we found partially slightly worse classifications, which implies that intention recognition in absence of overt behavior might be easier, but also that concurrent tasks/processes such as motor execution might (slightly) affect effects between differential decision outcomes. The arguably most closely related foregoing investigations required participants to press a key to indicate the intention to select (Bednarik et al., 2012; Medathati et al., 2020). The interaction of key presses might hamper predictions, possibly because two factors simultaneously affected physiological activation (rather than inhibition). Differentially large pupil sizes for targets relative to distractors, but constantly similar pupil sizes for distractors depending on stimulus probability, suggest an involvement of activation for targets rather than inhibition for distractors. When pressing a key, processes of activation are assumed, whereas the condition without key press can be seen as a variation of a NoGo condition that is generally associated with inhibition (Donders, 1868). Slightly better results under inhibition could thus result from effects of decision-making on pupil dilation being superimposed to a small degree by other factors

affecting activation or that inhibition accentuates effects. However, here reported differences are small and rather inconsistent, which is why such considerations should be read with caution; no interaction of key press and deciding was found using functional data analysis either (Strauch et al., 2020).

Our results demonstrate that classifiers that were trained on one fraction of data and tested on another fraction of data *within* participants did not outperform classifiers trained on data of some participants, but tested on data of others. Hence, the differential pupil response observed for targets relative to distractors did not depend on idiosyncratic features of participants. Hence, the differential pupil response observed for targets relative to distractors did not depend on idiosyncratic features of participants.

Besides the more standardized test environment, lower classification performances in the previous machine-learning investigations (Bednarik et al., 2012; Medathati et al., 2020) could be a direct result of the degree of consequence to a decision. In the presented experiment herein, participants were punished for errors; targets allowed to finish the experiment, which rendered decisions and targets in particular meaningful. Depending on the specific use-case, one might assume that self-initiated actions are substantially more meaningful than sometimes vague experimental instructions, since they reflect a person's inner goals. Better classification performances relative to previous investigations are even more striking given that we have run analyses multiple times, leading to presumable conservative results. As Bednarik et al. (2012), Kootstra et al. (2020), and Jang et al. (2014) demonstrate, the inclusion of further gaze characteristics can substantially improve classification in related settings, which should be addressed in an investigation that allows free gaze movements, but controls for effects such as foreshortening errors affecting pupil size.

A number of applications to our findings are conceivable for applications in research and technology, be it in understanding interpersonal communication (Einhäuser et al., 2010; Naber, Stoll, Einhäuser, & Carter, 2013; Quesque, Behrens, & Kret, 2019) or in allowing innovative human–computer interfaces (Strauch, Ehlers, & Huckauf, 2017).

Investigating visual search, pupil size could aid the inference of search targets in combination with gaze position (Spiller et al., 2021). In general, these results are most promising for contexts where overt responses are undesirable/impossible in research, for example, when motor execution might superimpose or convolute effects, such as in electroencephalography or functional magnetic resonance imaging research. Another field where intention recognition based on ocular features is of huge interest is human–robot interaction (Jang et al., 2014), where robots could use inferences to act proactively towards a human counterpart (Huang & Mutlu, 2016). For eyes-only interaction, although pupil size changes have been suggested to be informative of selection (Bednarik et al., 2012; Strauch, Ehlers, & Huckauf, 2017, Strauch, Greiter, & Huckauf, 2017) and could be used to enable basic communication with locked-in patients or patients with minimal conscious state (Stoll et al., 2013), the most common approach to identify users' intention to select a target stimulus remains gaze dwelling. That means that users need to keep their gaze position within a defined area for a predetermined time (dwell time) to select. Our results show that pupil dilation provides a promising candidate for a more natural eyes-only interaction that does not require users to look at a target for a prolonged, often unnatural, time or would at least allow reducing such times. Hereby, pupil-based intent classification could contribute to other intent-prediction mechanisms (Strauch, Huckauf, Krejtz, & Duchowski, 2018).

To conclude, using a dataset obtained from 69 participants, a random forest classification classifier was able to predict binary decision outcomes above chance (50%) with an AUC of up to 0.76 if targets were rare (25%) based on pupil dynamics alone. Predictions were gradually better, the rarer a target stimulus was, which is likely traceable to the oddball effect. But even decisions where targets were more frequent than distractors could be classified above chance level. Hereby, the first derivative of pupillary changes provided the most useful features and could be used as a measure in a variety of studies due to its timely and easy-to-calculate character. We see applications to this general phenomenon in research where intention cannot be communicated overtly via motor execution and applications such as gaze-based or assisted interaction, or human–robot interaction.

Keywords: pupillometry, classification, binary decision-making, oddball effect, machine-learning, random forest, pupil dilation

Acknowledgments

Commercial relationships: none.
Corresponding author: Christoph Strauch.
Email: c.strauch@uu.nl.
Address: Experimental Psychology, Helmholtz Institute, Utrecht University, Heidelberglaan 1, Utrecht, 3584 CS, the Netherlands.

References

Bednarik, R., Vrzakova, H., & Hradis, M. (2012). What do you want to do next: A novel approach

for intent prediction in gaze-based interaction. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications* (pp. 83–90), doi:10.1145/2168556.2168569.

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences of the United States of America, 111*(5), E618–E625, doi:10.1073/pnas.1317557111.

Donders, F. C. (1868). Over de snelheid van psychische processen. *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool (1868–1869), 2*, 92–120.

Eberhardt, L. V., Strauch, C., Hartmann, T. S., & Huckauf, A. (submitted). Increasing pupil size is associated with improved detection performance in the periphery.

Ehlers, J., & Meinecke, A. (2020). Voluntary pupil control in noisy environments. In *Proceedings of the 2020 Symposium on Eye-Tracking Research and Applications* (pp. 1–5).

Einhäuser, W., Koch, C., & Carter, O. L. (2010). Pupil dilation betrays the timing of decisions. *Frontiers in Human Neuroscience, 4*, 1–9, doi:10.3389/fnhum.2010.00018.

Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences, 105*(5), 1704–1709, doi:10.1073/pnas.0707727105.

Hakerem, G., & Sutton, S. (1966). Pupillary response at visual threshold. *Nature, 212*(5061), 485–486, doi:10.1038/212485a0.

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods, 48*(2), 510–527, doi:10.3758/s13428-015-0588-x.

Huang, C.-M., & Mutlu, B. (2016). Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 83–90), doi:10.1109/HRI.2016.7451737.

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science, 12*(4), 670–679, doi:10.1111/j.1467-7687.2008.00805.x.

Jang, Y.-M., Mallipeddi, R., Lee, S., Kwak, H.-W., & Lee, M. (2014). Human intention recognition based on eyeball movement pattern and pupil size variation. *Neurocomputing, 128*, 421–432, doi:10.1016/j.neucom.2013.08.008.

Jangraw, D. C., Wang, J., Lance, B. J., Chang, S.-F., & Sajda, P. (2014). Neurally and ocularly informed graph-based models for searching 3D environments. *Journal of Neural Engineering, 11*(4), 046003, doi:10.1088/1741-2560/11/4/046003.

Klingner, J. (2010). Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications* (pp. 275–282), doi:10.1145/1743666.1743732.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291–300, doi:10.1097/AUD.0b013e3182310019.

Kootstra, T., Teuwen, J., Goudsmit, J., Nijboer, T., Dodd, M., & Van der Stigchel, S. (2020). Machine learning-based classification of viewing behavior using a wide range of statistical oculomotor features. *Journal of Vision, 20*(9), 1–1, doi:10.1167/jov.20.9.1.

Martin, J. T., Whittaker, A. H., & Johnston, S. J. (2020). Component processes in free-viewing visual search: Insights from fixation-aligned pupillary response averaging. *Journal of Vision, 20*(7), 5–5, doi:10.1167/jov.20.7.5.

Medathati, N. V. K., Desai, R., & Hillis, J. (2020). Towards inferring cognitive state changes from pupil size variations in real world conditions. In *ACM Symposium on Eye Tracking Research and Applications* (pp. 1–10), doi:10.1145/3379155.3391319.

Murphy, P. R., O'connell, R. G., O'sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with bold activity in human locus coeruleus. *Human Brain Mapping, 35*(8), 4140–4154, doi:10.1002/hbm.22466.

Naber, M., & Murphy, P. (2020). Pupillometric investigation into the speed-accuracy trade-off in a visuo-motor aiming task. *Psychophysiology, 57*(3), e13499, doi:10.1111/psyp.13499.

Naber, M., Stoll, J., Einhäuser, W., & Carter, O. (2013). How to become a mentalist: reading decisions from a competitorâAZs pupil can be achieved without training but requires instruction. *PLoS One, 8*(8), e73302, doi:10.1371/journal.pone.0073302.

Pasquali, D., Aroyo, A. M., Gonzalez-Billandon, J., Rea, F., Sandini, G., & Sciutti, A. (2020). Your eyes never lie: A robot magician can tell if you are lying. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 392–394), doi:10.1145/3371382.3378253.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems* (pp. 5776–5788), doi:10.1145/2858036.2858117.

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision, 10*(10), 3, doi:10.1167/10.10.3.

Quesque, F., Behrens, F., & Kret, M. E. (2019). Pupils say more than a thousand words: Pupil size reflects how observed actions are interpreted. *Cognition, 190*, 93–98, doi:10.1016/j.cognition.2019.04.016.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology, 22*(2), 204–207, doi:10.1111/j.1469-8986.1985.tb01587.x.

Spiller, M., Ying-Hsang, L., Zakir, H., Gedeon, T., Geissler, J., & Nürnberger, A. (2021). Predicting visual search task success from eye gaze data as a basis for user-adaptive information visualization systems. *The ACM Transactions on Interactive Intelligent Systems*.

Stoll, J., Chatelle, C., Carter, O., Koch, C., Laureys, S., & Einhäuser, W. (2013). Pupil responses allow communication in locked-in syndrome patients. *Current Biology, 23*(15), R647–R648, doi:10.1016/j.cub.2013.06.011.

Strauch, C., Ehlers, J., & Huckauf, A. (2017). Pupil-assisted target selection (pats). In *Ifip Conference on Human-Computer Interaction* (pp. 297–312), doi:10.1007/978-3-319-67687-6_20.

Strauch, C., Greiter, L., & Huckauf, A. (2017). Towards pupil-assisted target selection in natural settings: Introducing an on-screen keyboard. In *Ifip Conference on Human-Computer Interaction* (pp. 534–543), doi:10.1007/978-3-319-67687-6_37.

Strauch, C., Greiter, L., & Huckauf, A. (2018). Pupil dilation but not microsaccade rate robustly reveals decision formation. *Scientific Reports, 8*(1), 1–9, doi:10.1038/s41598-018-31551-x.

Strauch, C., Huckauf, A., Krejtz, K., & Duchowski, A. T. (2018). Towards a selection mechanism integrating focal fixations, pupil size, and microsaccade dynamics. In *Eye Tracking for Spatial Research, Proceedings of the 3rd International Workshop* (pp. 9–15), doi:10.3929/ethz-b-000222294.

Strauch, C., Koniakowsky, I., & Huckauf, A. (2020). Decision making and oddball effects on pupil size: Evidence for a sequential process. *Journal of Cognition, 3*(1), 1–17, doi:10.5334/joc.96.

Van der Stoep, N., Van der Smagt, M., Notaro, C., Spock, Z., & Naber, M. (2021). The additive nature of the human multisensory evoked pupil response. *Scientific Reports, 11*(1), 1–12, doi:10.1038/s41598-020-80286-1.

van Kempen, J., Loughnane, G. M., Newman, D. P., Kelly, S. P., Thiele, A., O'Connell, R. G., . . . Bellgrove, M. A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *Elife, 8*, e42541, doi:10.7554/eLife.42541.

Wetzel, N., Einhäuser, W., & Widmann, A. (2020). Picture-evoked changes in pupil size predict learning success in children. *Journal of Experimental Child Psychology, 192*, 104787, doi:10.1016/j.jecp.2019.104787.

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences, 109*(22), 8456–8460, doi:10.1073/pnas.1201858109.