Learning from Peers’ Eye Movements in the Absence of Expert Guidance:
a Proof of Concept Using Laboratory Stock Trading, Eye Tracking, and Machine Learning
by
Michał Król² and Magdalena Król¹

¹ Department of Economics, School of Social Sciences, University of Manchester, Manchester, United Kingdom
² Wrocław Faculty of Psychology, SWPS University of Social Sciences and Humanities in Wrocław, Wrocław, Poland

* Corresponding author:
Email: michal.krol@manchester.ac.uk, The University of Manchester, School of Social Sciences, Oxford Road, Manchester, M13 9PL, UK

Keywords
debiasing; eye-tracking; machine learning; learning from peers; crowd-sourced process feedback

This work was supported by the Polish Ministry of Science and Higher Education grant number BST/WROC/2016/A/6.
Abstract

Existing research shows that people can improve their decision skills by learning what experts paid attention to when faced with the same problem. However, in domains like financial education, effective instruction requires frequent, personalized feedback given at the point of decision, which makes it time-consuming for experts to provide and thus prohibitively costly. We address this by demonstrating an automated feedback mechanism that allows amateur decision-makers to learn what information to attend to from one another, rather than from an expert. In the first experiment, eye-movements of N=100 subjects were recorded while they repeatedly performed a standard behavioral finance investment task. Consistent with previous studies, we found that a significant proportion of subjects were affected by decision bias. In the second experiment, a different group of N=100 subjects faced the same task but, after each choice, they received individual, machine-learning-generated feedback on whether their pre-decision eye-movements resembled those made by Experiment 1 subjects prior to good decisions. As a result, Experiment 2 subjects learned to analyze information similarly to their successful peers, which in turn reduced their decision bias. Furthermore, subjects with low Cognitive Reflection Test scores gained more from the proposed form of process feedback than from standard behavioral feedback based on decision outcomes.
1 Introduction

Previous research on ‘Eye Movement Modeling Examples’ (EMME) demonstrated that learners can benefit from seeing a visualization of the eye-movements of an expert explaining the task at hand (Jarodzka et al., 2012; Jarodzka, van Gog, Dorr, Scheiter, & Gerjets, 2013). Apart from directly superimposing the experts’ eye-movements on instruction material, similar gains are achieved by using visual cues based on experts’ observed attentional patterns (Thomas & Lleras, 2007). For example, trainee doctors could learn to better diagnose illnesses after seeing examples of medical images with highlights indicating which parts of the image expert doctors fixated on before reaching a certain conclusion. However, in some procedural problem-solving tasks, EMME proved less effective (van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009; van Marlen, van Wermeskerken, Jarodzka, & van Gog, 2016).

In this paper, we addressed three problems that we believe limit the applicability of eye movement modeling. First, a model of how the task should ideally be solved could at times be costly and difficult to obtain, e.g. due to sufficiently competent experts being unavailable or too few (see Gegenfurtner et al., 2017 for a discussion of this issue). Second, learners might differ in their characteristics, so showing everyone the same eye movement model could be less effective than customized instruction. Third, recent findings suggest that learners struggle to self-monitor their eye-movements (Kok, Aizenman, Võ, & Wolfe, 2017), and may, therefore, be unaware that they do not follow the ideal gaze model that they have seen.

As an example, a fundamental issue in financial education is that finance experts are generally unable to consistently provide guidance of high enough quality to justify their substantial professional fees (Bessler, Blake, Lückoff, & Tonks, 2017; Guercio & Reuter, 2014). In addition, amateur investors’ responsiveness to instruction may be heterogeneous, determined
by their varying susceptibility to decision bias (Dhar & Zhu, 2006) and other cognitive constraints (Carpena, Cole, Shapiro, & Zia, 2017). Any increases in financial literacy that do result from education campaigns tend to be temporary, unless the newly gained knowledge is immediately put into practice (Fernandes, Lynch, & Netemeyer, 2014). Thus, as argued by Willis (2011), effective financial education should be decision-specific rather than generic, provided regularly at the point of decision-making, and tailored to the individual learner's needs. Combined with the high cost of financial experts’ time, all of this makes such an education prohibitively costly.

We therefore designed a modification of the EMME technique, whereby amateur decision-makers can learn from their peers in the absence of experts, and tested the new approach in the context of financial decisions.

Specifically, we conducted two laboratory stock market experiments, based on a well-established behavioral finance framework (used by Fischbacher, Hoffmann, & Schudy, 2017; Frydman, Barberis, Camerer, Bossaerts, & Rangel, 2014; Frydman & Camerer, 2016; Kuhnen, 2015, among others). Subjects trade stocks, each of which can be either ‘good’ (more likely to increase in price than to decrease) or ‘bad’. Although they are not aware of whether a given stock is good, they can infer this from its observed price path. The abovementioned existing studies reported that, faced with this problem, people typically exhibit a bias to realize gains rather than losses, known as the disposition effect (Odean, 1998). This leads to bad decisions, because a stock that registered gains is likely to be a good stock and should therefore not be sold, whereas one that registered losses is likely bad and should be disposed of (that is, one should realize losses rather than gains).
In fact, the disposition effect is of considerable practical importance as one of the main reasons for the departures from the efficient market hypothesis observed in financial markets. The underlying mechanism is that biased investors tend to ignore evidence that a stock they own is bad and as such likely to subsequently decrease (Frazzini, 2006). Thus, incoming information is not fully priced in straight away, creating short-term autocorrelation of the price changes (known as drift or momentum). The more sophisticated institutional investors can exploit the drift by selling (or, indeed, short selling) the stock before their biased individual counterparts’ delayed response brings prices further down (Ke & Ramalingegowda, 2005). In other words, the experimental framework we use here captures an important real-world problem, and one where no effective debiasing solutions currently exist.

Accordingly, the purpose of our first experiment, in which subjects did not receive any feedback, was to train and validate a machine-learning algorithm that can predict, based on a person’s eye-movements in a given decision trial, whether or not she will make a good decision (to sell or not to buy a stock that is most likely bad, or to buy or not to sell one that is most likely good). In so doing, we build on recent studies that used machine-learning to identify the objectives pursued by subjects (Borji & Itti, 2014) or their decision strategies (Król & Król, 2017). Here, we hypothesized that, based not only on what pieces of information our subjects look at, but also on the order in which they do so, it should be possible to detect if a person’s decision-making process is sound and to predict if a good choice will follow.

In the second experiment, designed based on the results of the first study, a different group of subjects were each given ‘eye-feedback’ after every trading decision, generated by the previously trained machine-learning algorithm. That is, they were given ‘good’ (that is, positive) feedback when their eye-movements in a given decision trial were similar to those made by
Experiment 1 subjects prior to good decisions (and ‘bad’, i.e. negative feedback in the opposite case). Existing research on EMME shows that novices are more likely than experts to become distracted by task-irrelevant information (Wolff, Jarodzka, van den Bogert, & Boshuizen, 2016).

Thus, we hypothesized that the new form of feedback proposed here would be especially helpful (in terms of debiasing decisions) to people whose individual traits make them prone to automatic reactions based on irrelevant data, and to cognitive bias in general.

To verify this idea, we asked Experiment 2 subjects to complete the Cognitive Reflection Test (CRT, see Frederick, 2005), a widely used tool measuring susceptibility to heuristics and biases, including the ‘conservatism bias’ to learn insufficiently from new information (Hoppe & Kusterer, 2011; Oechssler, Roider, & Schmitz, 2009). To see how cognitive reflection determines the effectiveness of the proposed form of eye-feedback relative to more standard forms of decision support, a separate group of Experiment 2 subjects received ‘behavioral’ feedback, where after each choice they are simply told if their decision was good or not. Thus, our two types of feedback mirrored the classic distinction between ‘process’ and ‘outcome’ feedback (Earley, Northcraft, Lee, & Lituchy, 1990), and the comparison between them contributes to the debate on the extent to which process feedback on cognitive operations employed to solve problems is essential for self-regulated learning (Winne, 2005). Although we have evaluated the proposed technique in the context of a specific decision problem, in the General Discussion section we argue that the results could shed light on interesting general problems with potentially wide-ranging applications, particularly on the feasibility of creating crowdsourced debiasing techniques supported by artificial intelligence.
2 Experiment 1

2.1 Method

2.1.1 Subjects.

One hundred student subjects were recruited at a large private university, excluding students of economics or finance. Although no studies considered an effect similar to ours before, making power analysis difficult, we chose a number of subjects larger than in most prior studies based on the current experimental stock market design (e.g. Frydman & Rangel, 2014 featured 58 subjects). Their average age was 22.3 (SD=3.59), 56 of them were female, and all had normal or corrected to normal eyesight. Both studies were approved by the local Faculty Research Ethics Committee and conducted in accordance with the Helsinki Declaration.

2.1.2 Stimuli and Design

Experimental Stock Market. Following the protocol of a number of previously mentioned behavioral finance studies, subjects were given the opportunity to trade three stocks. Initially, each subject received 200 units of experimental currency (‘points’), part of which was exchanged to one unit of each stock at an initial price of 50. In each of 80 decision trials, a randomly chosen stock was subject to a price update, after which the subject would decide whether to buy or sell a unit of that stock, thus completing the trial.

The price path of each stock was governed by a two-state Markov chain with a good state and a bad state. If the stock subject to a price update is in the good state then its price increases by one with probability 2/3 and decreases by one with probability 1/3. Conversely, if the stock is in the bad state, its price increases with probability 1/3 and decreases with probability 2/3. Initially, the state of each stock was independently drawn as good or bad with equal probability,
and subjects would not observe the true states but could infer them from the observed price paths.

In addition, whenever a stock undergoes a price update, there is a 20% chance that its state will change from good to bad or vice-versa. The price path was randomized independently for each subject, since we wanted to test a method of using eye-data of people who traded stocks in the past to give feedback to other people in different market circumstances.

To illustrate how one could infer the state of a stock, suppose that a price change $z_n \in \{-1,1\}$ ensued during its n-th price update, and that the probability that it was in a good state at its previous price update was $p_{n-1}$. The probability that the stock is currently in a good state can then be recursively calculated as:

$$p_n(z_n, p_{n-1}) = \frac{(0.5 + z_n/6)(0.6p_{n-1} + 0.2)}{(0.5 + z_n/6)(0.6p_{n-1} + 0.2) + (0.5 - z_n/6)(0.8 - 0.6p_{n-1})} \quad (1)$$

(prior to the first price update we have $p_0 = 0.5$).

Put simply, if the price of a stock is seen to go up, it becomes more likely that it is a good stock, and therefore that its price will increase again in the future (for a more detailed derivation and explanation of the above formula, see e.g. section II.B of Frydman et al., 2014).

The optimal trading strategy of a risk-neutral Bayesian investor is therefore: a) to buy an asset whenever $p_n > 0.5$, i.e. when it is most likely in a good state, and hence most likely to increase at the next price update; b) to sell an asset when $p_n < 0.5$, i.e. when it is most likely in a bad state, and so most likely to decrease at the next price update. We will henceforth refer to observed decisions consistent with these rules as ‘good’, and to all other decisions as ‘bad’. The definition of good decisions assumes risk-neutrality, following the aforementioned existing studies that use the same setting as we do here, and the same definition. Still, one could, in
principle, specify an alternative one, incorporating risk-aversion, in which the buy/sell threshold is above 0.5. Despite opting for the risk-neutral threshold as a natural, ‘focal’ choice, upon collecting the data we verified that slight variations in the threshold do not cause a significant change in the behaviour of the developed feedback mechanism (see below). What matters for its training is that a rational trader, irrespective of the degree of risk aversion, should be more likely to buy or hold a stock as \( p_n \) increases.

*Decision Screen.* An example decision screen (always preceded by a centrally-located fixation cross) is shown in Fig. 1. The icon at the top of the decision screen represents the asset subject to a price update: we use square, circle, and triangle to represent the three stocks. A smaller ‘wallet’ icon superimposed on the shape indicates that the stock is currently owned by the subject (otherwise the interior of the shape is empty). Below, on the left, a sequence of five icons shows the history of recent price changes of the stock (most recent ones on the right). An up arrow represents a positive (+1) change in the price, and a down arrow represents a negative (-1) change. On the right, the current price of the stock is shown, together with either: a) the price at which it was purchased, if the subject currently owns the stock; or b) if the subject does not own the stock, the price at which it was most recently sold. In case a), the two choice buttons at the bottom of the decision screen represent the option to either sell the stock or not, while in case b) the two options (choice buttons) are to either buy the stock or not.

Every time one decided to buy a stock that is not already owned, the current price of the stock would be subtracted from the pool of uninvested experimental currency (points) and the stock added to the portfolio. When selling a previously bought stock, the opposite process occurred (there were no transaction costs). Subjects could trade each of the three assets any number of times (whenever its price was subject to update).
Figure 1. An example decision screen (red elements are for information and were not seen by subjects). In this example, the ‘circle’ stock is owned by the subject (as indicated by the ‘wallet’ icon), who may therefore sell it or hold. The price of the stock went up during the latest price update, and was up three times in the last five price updates.
Portfolio Information Screen and Payoffs. In every trial, once the subject made the decision as explained above, the decision screen was replaced by a ‘portfolio information screen’, depicted in Fig. 2. This showed the current prices of the owned stocks, the amount of uninvested experimental currency (‘points’) and the resulting ‘total value’ of the portfolio. It also displayed the monetary payoff the subject would have received if the experiment were to end at that moment. This was calculated by subtracting the initial points’ allocation (200) from the current total value, multiplying this difference by the equivalent of 0.40 USD in the experiment’s local currency, and adding an equivalent of 8 USD show-up fee to the result (the study took approximately 35 minutes to complete).
Figure 2. An example portfolio information screen shown to subjects after they submit their decision. Here, the ‘circle’ asset is not currently owned (no wallet icon inside the shape), hence its value in the portfolio is zero.
The stimulus presentation software was programmed in Wolfram Mathematica. Each subject was tested individually, seated at a laptop operating Microsoft Windows 8, with a 15.4-inch screen with the resolution set to 1280x720, and a SensoMotoric Instruments RED250 mobile eye tracking device attached underneath the screen and set to 250Hz frequency. At the beginning of the study, once the subject has read the on-screen instructions, we conducted a standard five-point semi-automatic calibration procedure (the average deviation was below 0.5° for all subjects). A headrest was used to stabilize the head position and ensure that the distance between subject's eyes and the device was approximately 65cm. To detect eye-fixations, we used the SMI high-speed fixation detection algorithm with required fixation duration 80ms and max. dispersion 100px.

During the first ten decision trials, the subjects could familiarize themselves with the task, taking as much time as needed to submit choices. From then on (following an instruction slide warning of the change), there was a time limit of 13.5 seconds per trial initially, decreasing by 1.5 seconds in each of the subsequent five trials, and fixed at 6 seconds from the sixteenth trial onwards (this was slightly more than in existing studies based on the same experimental stock market protocol). After the time limit has passed, there would be a sound signal reminding the subject to enter the choice, and failing that, a further 3 seconds later the choice would be made at random by the experimental software (this happened in less than 1% of trials).

In this section, we provide a technical description of the training and validation procedure of the machine-learning algorithm assessing one’s decision-making based on eye-data.
Preliminary Data Processing. From the sixteenth decision trial onwards, i.e. since the final time limit was in place, we collected each subject's eye-fixations which occurred while the decision screen was shown, dropping trials in which the subject failed to reach a decision within the time limit.

For each fixation, processed in chronological order, we then determined which of the areas of the decision screen it was located in. Specifically, using eye tracking terminology, we distinguish between four ‘Areas-of-Interest’ (AOIs): (1) the price history panel, comprising the five arrows representing the five most recent price changes of the stock; (2) the box showing the purchase price or the most recent selling price (for stocks not currently owned); (3) the box showing the current price; (4) the two available choice buttons, treated as a single AOI. We dropped fixations that were not located in any AOI. Thus, for each decision trial, we obtained a sequence of integer numbers between 1 and 4 (‘scan path sequence’), representing the order in which the subject looked at elements of the decision screen.

Next, we extracted statistical regularities from the scan path sequence via the temporal-difference learning algorithm described and tested in Hayes, Petrov, and Sederberg (2011), who used it to predict subjects’ cognitive capacity and intelligence, a similar problem to estimating the ‘capacity to make good decisions’, which is what we pursue here. The algorithm incrementally builds a 4x4 Scanpath Successor Representation (SR) matrix, representing the temporally discounted number of expected future visits to all AOIs (rows), given the subject fixated on any individual AOI (column). Importantly, the SR matrix encodes not only direct but also longer-term transitions between the AOIs. For instance, given a scan path of 1,2,4,1,2, the algorithm would encode not only the transition between 1 and 2 but also the longer-term transition between 1 and 4. The longer-term transitions are discounted at a factor of $\gamma \in (0,1)$. 

14
while another parameter, \( \alpha \in (0,1) \) determines the learning rate. The optimal parameter values were found in the cross-validation process described below.

Apart from the SR matrix, for each decision trial, we also computed, using formula (1), the probability \( p_n \) of the stock currently being in the good state. We associated the SR matrix with an index of ‘decision goodness’, ranging from 0 to 1, and given by \( p_n \) for decisions to buy or hold a stock, or \( 1 - p_n \) for decisions to sell or not buy a stock. For example, suppose that one sells a stock which, taking available evidence into account via the Bayesian posterior probability formula (1), has a 60% chance of being in the good state. Then the decision goodness is equal to \( 1 - 0.6 = 0.4 \). But if the stock had a 40% probability of being in the good state, then the decision goodness would equal 0.6. Naturally, an index of decision goodness below 0.5 corresponds to what we have earlier defined as a bad decision, and one above 0.5 corresponds to a good one.

Decision goodness is, therefore, a continuous version of the good/bad dichotomy, recognizing that some good decisions are better than others, etc. Put simply, the more likely the stock is to be in the good state, the better any decision to buy or hold it, and the worse any decision to sell or not buy it. The purpose of using decision goodness instead of a binary index is to increase precision and reduce noise, because a choice to sell a stock with a 51% probability of being in the good state is actually not an extremely bad decision and should not be considered on par with one to sell a stock with an 80% chance of being in the good state.

**Predicting Decision Goodness.** We aimed to predict the decision goodness of a given decision trial, by comparing the subject’s eye movements during the trial encoded in the SR matrix with a training set of other subjects’ encoded eye-movements. We did this separately for each type of decision, i.e. for decisions on whether to sell vs. those on whether to buy a stock. Separating the two types of decisions was motivated by the fact that existing literature suggests
that very different forms of decision bias are at work in each case, with those affecting the selling
decisions having a more significant effect on behavior (see e.g. Frydman & Camerer, 2016;
Strahilevitz, Odean, & Barber, 2011; we will revisit this issue when examining the results). Thus,
the same eye-movement patterns that might be associated with bad selling decisions could mean
something quite different in the context of choosing whether or not to buy a stock.

The first step in predicting decision goodness was using principal component analysis to
project the $4 \times 4 = 16$ dimensional SR data to an $N < 16$ dimensional approximating manifold.
Next, we used the Partitioning-around-Medoids algorithm (PAM) to cluster the dimension-
reduced SR's in the training set into $K$ clusters. PAM selects a set of $K$ representative points
(‘medoids’) to form initial clusters, with each of the other data points assigned to its nearest
medoid. The algorithm then iteratively updates the set of medoids to reduce the distance between
them and the rest of the data.

With the training set partitioned in this way, for any dimension-reduced SR matrix outside the
training set, we can calculate a predicted decision goodness as a sum of the average values of
decision goodness in each cluster weighted by inverse distances between the SR in question and
the clusters’ medoids. In other words, if the SR is similar to clusters generally associated with
good decisions, then the predicted decision goodness is going to be high as well$^1$.

Thus, the procedure is similar to the one used by Borji and Itti (2014), who showed that it
possible to predict, based on eye-data, which of a number of image-viewing tasks (such as
estimating the material circumstances of people shown in the image) a person performs. The
difference is that we use a clustering algorithm instead of the (closely related) ‘nearest

---

$^1$ To implement the outlined procedure, we used the Wolfram Research Mathematica 11.3 software package,
particularly its ‘Cluster Classify’ and ‘Dimension Reduction’ algorithms, adding a custom-coded implementation
of the Scanpath Successor Representation technique. Code snippets illustrating how the classification algorithm is trained
to evaluate unseen scanpaths are available on the project’s Open Science Framework page (https://goo.gl/3E3wrC).
neighbors’ technique. This is because the former method is recommended when, as in our case, the data is likely to exhibit a problem known as ‘class label noise’ (Frénay & Verleysen, 2014). Specifically, a poor decision-maker could still make a good trade by accident or ‘for the wrong reasons’. For example, suppose that one has held the stock for a long time, incurring large losses, but that only recently the stock has started to make gains, suggesting that it has switched to the good state. The investor does, indeed, choose to continue to hold, but not because she is acting according to the formula (1), but because, driven by the disposition effect, she is reluctant to realize what is still an overall loss. Conversely, a bad decision could be made in a borderline case by a person whose reasoning is generally sound. Either way, the accompanying eye-data representation will be mislabeled, i.e. an example of bad decision-making will be labeled as good, or vice-versa.

We conducted cross-validation to evaluate prediction performance and optimize parameter values through grid search, where $\alpha, \gamma \in \{0.1, 0.3, \ldots, 0.9\}$, $K \in \{10, 20, \ldots, 50\}$, and $N \in \{1, 2, 4, 8, 16\}$. Specifically, for each type of decision, each parameter combination and each of the 100 subjects, we assigned the remaining 99 subjects’ decision trials of the given type to the training set, and predicted the decision goodness of each of the singled-out subjects’ decisions of the same type using the given parameter combination. Having done the same for each subject, we recorded the mean squared error between the predicted and actual values of decision goodness, and proceeded to the next parameter combination, eventually choosing the one minimizing the error.
2.2 Results and Discussion

**Behavioral Indicators of Decision Bias.** We calculated the ‘proportion of gains realized’ (PGR) in the sense of Odean (1998) for all decisions on whether or not to sell an owned stock (henceforth, ‘selling decisions’). This is defined as the share of trials in which a stock trading at a gain relative to purchase price is sold in the total number of trials in which an owned stock is trading at a gain. The ‘proportion of losses realized’ (PLR) is defined analogously. Across all subjects and trials, we found PGR = 45.0% and PLR = 22.7%. A sign test indicated that the PGR - PLR difference (calculated separately for each subject, M=26.5%) was greater than zero, Z=-4.671, p<.001. Accordingly, the proportion of good (Bayesian-optimal) selling decisions (M=45.6%) was smaller than the ‘random choice’ level of 50%, Z=-2.96, p=.003. Thus, the results indicate a significant ‘disposition effect’ (tendency to realize gains rather than losses), as reported by previous research using the same experimental setting (e.g. Frydman et al., 2014).

Analogously, for buying decisions, we found that across all subjects/trials the ‘proportion of stocks up since being sold that are repurchased’ (PUR) in the sense of (Strahilevitz et al., 2011) is 35.8%, while the ‘proportion of stocks down since being sold that are repurchased’ (PDR) is 39.8%. The difference is used to measure the ‘repurchase effect’, which mirrors the disposition effect, in that subjects avoid repurchasing an asset trading at a gain relative to the price at which they previously sold it. A paired sign test indicated that the PDR - PUR difference (M=5.8%) was lower than PGR - PLR (Z=-2.34, p=.025), and a one-sample sign test showed that it was not significantly above zero, Z=-.919, p=.358. Additionally, a paired signed-rank test indicated that the proportion of good buying decisions (M=51.6%) was higher than that of selling decisions (M=45.6%), W=3118.5, p=.014. Thus, the influence of the disposition effect on
subjects’ selling decisions was strong, possibly stronger than that of the repurchase effect on buying decisions.

**Configuring and Evaluating the Machine Learning Algorithm.** Regarding our main objective of using SR data to predict decision goodness, the optimal parameter configuration found in the cross-validation procedure was $\alpha = 0.3, \gamma = 0.5$ for buying decisions, $\alpha = 0.1, \gamma = 0.9$ for selling decisions, and $N=2, K=40$ for both types of decisions (the proportion of total SR variance explained by a principal components reduction from 16 to $N = 2$ dimensions was 53.55% for selling decisions and 45.17% for buying decisions). We used these parameter values and the entire set of Experiment 1 data to train the algorithm that will give feedback to subjects in Experiment 2. To make the feedback easy for subjects to process, we decided to communicate it as either ‘good’ or ‘bad’ (instead of stating the exact predicted decision goodness).

Thus, the question was where to set the predicted decision goodness threshold above which the subject will receive good feedback, which is effectively a signal detection problem (where we wish to detect good decisions and reinforce them with good feedback). If we set the threshold low, then the ‘sensitivity’ or ‘true positive rate’ of the feedback mechanism would be high, i.e. a person who makes a good decision would be likely to receive good feedback. However, this would come at a cost of a high ‘false alarm’ or ‘false positive’ rate, i.e. a person who makes a bad decision would also be likely to be reinforced with good feedback. To find the threshold that would optimally balance these two metrics, we used Experiment 1 data to construct the ‘receiver operating characteristic curve’ (ROC) combining all possible false alarm-sensitivity pairs (points) that would obtain depending on where the threshold is set, based on the values of predicted decision goodness obtained in the cross-validation process for the optimal parameter combinations. This is depicted in Fig. 3.
The figure also includes, as a benchmark for our primary model, a second ROC that would result from using a ‘simple’ prediction algorithm based on fixation times, specifically one in which good feedback is given if the proportion of trial duration the subject spent looking at AOIs 2 and 3 (past/current prices) was below a given threshold.

The ‘area under the curve’ (AUC) of the primary model (which measures its overall accuracy across all threshold values), equal to 0.62, was significantly greater than the AUC of the benchmark model, equal to 0.59, based on DeLong’s test for two correlated ROC curves, Z=4.04, p<.001. Both AUC’s were significantly greater than the 0.5 chance level (p<.001), represented by the 45° line. The value of the predicted decision goodness threshold maximizing the difference between sensitivity and false alarm rate was 0.502, resulting in sensitivity close to 40% and a 20% false alarm rate. The cross-validated accuracy rate of the primary model corresponding to the optimal threshold under the optimal parameter configuration was 59.6% for buying decisions and 63.1% for selling decisions, and binomial tests indicated that both of these values were significantly above chance (p<.001).
Figure 3. The receiver operating characteristic curve for both types of decisions under the optimal parameter configuration in the primary prediction model (black) and a benchmark model based on fixation times (gray). The threshold values on the right vertical axis correspond to the black ROC, and the dashed 45° line represents the chance level.
Thus, despite the fact that we predicted each subject’s decision goodness based on data from other subjects, the prediction accuracy was greater than chance, and there were gains from considering the order in which information was processed, rather than just the spatial distribution of attention (as in the benchmark model).

Properties of the Machine Learning Algorithm. To see precisely how our primary model used scan path representation data to make predictions, consider a contour plot of predicted decision goodness of selling decisions across the range of values of the N=2 principal components of the data, shown in the top half of Fig. 4. Generally speaking, higher values of both component scores were associated with higher predicted decision-goodness (as illustrated by more ‘warm’ colors to the north-east of the graph).
Figure 4. The predicted decision goodness of selling decisions depending on eye-transitions made by the subject. In the top panel, a contour plot of decision goodness is shown, across the range of the two-dimensional approximating manifold that the 4x4 = 16 dimensional SR data was projected to via principal component analysis (based on the optimal parameter values derived in the cross-validation process). Each possible scanpath maps (via a linear transformation of the SR matrix) to a point in the graph, its coordinates being the N = 2 principal component scores. Warmer colors indicate higher predicted decision-goodness, the dotted line separates the range of values where it exceeds 0.5, while 95% of Experiment 1 data was contained within the region enclosed by the thick dashed line. The structure of the two principal components (i.e. the way in which a scanpath is mapped onto a point) is shown in the two panels underneath. The row/column numbers of the two weight matrices (one for each principal component/axis of the top graph) represent the AOIs (recall AOI 1 = price change history; 2 = purchase/selling price; 3 = current price; 4 = choice buttons). A warmer color (larger positive weight) indicates that transitions between the two given row/column AOIs increase the component score to a greater extent; cold colors (negative weights) mean that the component score is instead reduced by these transitions.
To understand the positive relationship between both component scores and predicted decision-goodness, consider the weight matrices of the two components depicted in the bottom half of Fig. 4. The first component (X-axis of the contour plot) penalizes repeated fixations on the price level information (AOI’s 2 and 3), or re-visiting this information once it had already been seen before. This is probably due to the fact that price level data was less relevant or useful for making Bayesian-optimal choices compared with the price change history (AOI 1). Hence, considering this information under time pressure at the very least left the subject with less time to analyze the key price change history data. In addition, comparing the current and purchase prices (AOI’s 3 and 2) is a pre-requisite for calculating the capital gain and hence for the disposition effect to influence decisions. Thus, transitions between those AOIs may indicate bias. This is in line with Frydman and Rangel (2014), where hiding purchase price information reduced the disposition effect.

The second component (Y-axis of the contour plot) rewards transitions to and from the price change history (AOI 1), particularly to the choice buttons (AOI 4) and from price level information and choice buttons back to AOI 1. Thus, a good decision-maker might still perform a cursory check of price level information, but would not fixate on it repeatedly or re-check it. Such a person would instead return to the price change history data and possibly proceed to the choice buttons directly from there, before finalizing the choice. This is in line with research showing that good performance in repeated tasks relies on executing special purpose visual routines that process only the minimum required information (Hayhoe, 2000), and that this allows for faster learning with fewer errors (Hullinger, Kruschke, & Todd, 2015). Consistent with existing work in behavioral finance (Chang, Solomon, & Westerfield, 2016; Kuhnen, Rudorf, & Weber, 2017), focusing on the price change history directly before making the choice...
could also indicate a lower cognitive dissonance or more accurate updating of one’s beliefs, no longer distorted by information about one’s subjective experience with the stock (also see Lejarraga, Woike, & Hertwig, 2016 for more work on learning from experience vs. description in experimental stock markets).

The above discussion applies equally to data from buying decision trials, with good decision-making associated with high values of two similarly structured principal components. Variation in Algorithm Accuracy across Subjects. It is also interesting to consider how the predicted decision goodness, as well as the actual capacity to make good decisions, are distributed across subjects. In particular, we found that the proportion of good decisions a subject has made was positively correlated with the average predicted decision goodness for that subject. This holds both for the buying decisions (Spearman’s $\rho=.593, p<.001$) and for selling decisions ($\rho=.532, p<.001$). Relatedly, Fig. 5 shows that practically all subjects whose predicted decision goodness was above the 0.502 threshold most of the time (points to the right of the dashed vertical line) also made good decisions in the majority of the trials. However, nearly a third of subjects whose predicted decision goodness was below the threshold also made good decisions in the majority of the trials. Those subjects performed relatively well despite looking at the past and current price information a lot, i.e. were apparently immune to any distortion in reasoning that this data induced in other subjects. If assessed by the trained feedback mechanism, they would have received bad feedback most of the time, which might have led to an unnecessary change of a sound decision-making strategy.
Figure 5. The median predicted decision goodness and the fraction of good decisions (pooled data for selling and buying) calculated for all subjects, each represented by a single point.
Implications for Experiment 2. Consequently, prior to running the second experiment, we expected that the effectiveness of eye-tracking feedback might vary depending on the subjects’ individual traits, and that it may be especially beneficial for subjects characterized by inability to suppress biased heuristic responses and reflect on the incoming information in an undistorted manner. This motivated our use of the Cognitive Reflection Test (which measures those abilities) for all subjects taking part in Experiment 2. Additionally, we expected eye-tracking feedback to have an impact on selling rather than buying decisions, as in the latter case we did not find significant evidence of decision bias.

All in all, we hypothesized that, in case of selling decisions: a) subjects who receive eye-tracking feedback will adjust their eye-movements so that predicted decision goodness will increase compared with Experiment 1; b) the change in eye-movement patterns will translate into a greater proportion of good decisions and lower measures of decision bias compared with Experiment 1; and c) the previous effect (b) will be particularly strong in subjects with low CRT scores, to the extent that these subjects may perform better under eye-tracking feedback than under a more standard form of feedback based on decision outcomes.
3 Experiment 2

3.1 Method

3.1.1 Subjects

We recruited a second group of 100 subjects from the same student population (mean age 22.8, SD =3.25, including 59 female subjects), excluding subjects who took part in Experiment 1.

3.1.2 Procedure

The study was identical to Experiment 1, with the following two exceptions. First, after every decision trial from the sixteenth trial onwards, i.e. since the final time limit was imposed, subjects received one of two types of feedback (described below). Each subject received the same type of feedback in all trials, i.e. the two types of feedback constituted two between-subjects experimental conditions. Second, before taking part in the study, each subject completed the Cognitive Reflection Test (described below).

3.1.3 Stimuli and Design

*Behavioral feedback.* In this type of feedback, after the subject finalizes the choice, but before the decision screen is replaced by the portfolio information screen, the elements of the decision screen are recolored for three seconds depending on whether or not the choice was consistent with that of a risk-neutral Bayesian investor. In particular, if the subject chose to sell or not to buy a stock when it is most likely to subsequently increase \( p_n > 0.5 \), or chose to hold or buy a stock when it is most likely to decrease \( p_n < 0.5 \), then the decision screen is recolored in red (bad feedback), as shown in Fig. 6. Otherwise, the screen is similarly re-colored in green.

This is also how we explained the feedback to subjects, except that we did not refer to \( p_n \) (see the attached appendix file for exact instructions given to subjects).
Figure 6. An example feedback screen shown to subjects in stage two (here, the screen is colored in red, indicating ‘bad’ feedback after a decision to sell was made).
Eye Tracking Feedback. The difference between this and the behavioral feedback described above (i.e. the other experimental condition) was that whether good (green) or bad (red) feedback is shown depends exclusively on the subject’s eye-movement patterns during the current decision trial, assessed using the algorithm trained on Experiment 1 data. Specifically, after each trial, the subject’s eye-fixations are collected, converted to a scan path sequence, then to a scan path successor representation matrix (using the optimal parameter configuration) and finally to a point by means of the previously derived dimension reduction function. This is then compared to the medoids of Experiment 1 data clusters, and the predicted index of decision-goodness is calculated. If the calculated index is below the optimal threshold of 0.502, the subject receives bad (red) feedback as described above for behavioral feedback. Otherwise, the feedback is good (green).

We explained the eye-feedback to subjects in a similar way to behavioral feedback. Specifically, we told subjects that the screen will be colored red if they visually processed information similarly to other subjects prior to choosing to sell or not to buy a stock that is most likely to subsequently increase, or prior to choosing to hold or buy a stock that is most likely to decrease. Otherwise, the screen would be colored green (again, see the Appendix for exact instructions).

Cognitive Reflection Test. The test is a task designed to measure one’s ability to override incorrect intuitive responses to problems (see Frederick, 2005 for a detailed description). In a variety of heuristics and biases tasks, the test has been shown to ‘predict rational-thinking performance independent not only of intelligence, but also of executive functioning and thinking dispositions, [accounting for] more unique variance explained than the block of intelligence measures’ (Toplak, West, & Stanovich, 2011). To illustrate, the first of the three questions in the
test is: ‘A bat and a ball cost 1.10$ in total. The bat costs 1.00$ more than the ball. How much
does the ball cost?’. The intuitive heuristic answer is 0.10$, but subjects able to suppress it would
provide the correct ‘reflective’ answer of 0.05$ (0.05$ + 1.00$ + 0.05$ = 1.10$). A correct
answer to each question is worth one point. Interestingly, despite the fact that it comprises only
three questions, the test has recently been shown to be highly robust to repeated exposure (Bialek
& Pennycook, 2017).

Allocation of Subjects to Conditions. We randomly assigned subjects to the two feedback
conditions (eye-tracking and behavioral feedback), with half of those who received a given test
score (0,1,2, or 3) assigned to each condition. This was achieved by assigning the first subject
who achieved a given score to a randomly chosen condition, and then assigning each subsequent
subject with the same score to a different condition than the previous subject who achieved that
score.

3.2 Results

3.2.1 CRT Scores

The number of subjects who received CRT scores of 0,1,2 and 3 was 28,24,25 and 23
respectively, i.e. a distribution similar to that obtained by Frederick (2005) for students of US
universities such as the Carnegie Mellon University (higher scores were obtained at Princeton or
MIT). As a result, 24 subjects with a ‘high’ score (2 or 3, henceforth, ‘reflective’ subjects as per
the dichotomy introduced by Cueva et al., 2016) and 26 subjects with a ‘low’ CRT score (0 or 1,
henceforth, ‘irreflective’ subjects), were assigned to each of the two feedback conditions. Thus,
the number of subjects per cell, while relatively small, exceeded the minimum of 20 subjects per
cell recommended by Simmons, Nelson, and Simonsohn (2011), and was similar to the numbers
seen in closely related previous studies (e.g. Frydman & Rangel, 2014).
3.2.2 Predicted Decision Goodness

The average predicted goodness (calculated by the machine learning algorithm based on eye-movements, and averaged across each subject’s decision trials from introducing the feedback onwards) of selling decisions under eye-feedback (M=.499), was higher than in Experiment 1 (M=.483), Mann-Whitney U=1743.5, p=.004. Similarly, the predicted goodness of selling decisions under behavioral feedback (M=.489) was higher than in Experiment 1, U=1914, p=.029.

As for the buying decisions, their predicted goodness under eye tracking feedback (M=.524) was higher than in Experiment 1 (M=.512), U=1795.5, p=.010, but the predicted goodness under behavioral feedback (M=.515) was not significantly higher than in Experiment 1, U=1958.5, p=.056.

For selling decisions, there was a significant positive correlation between predicted goodness and the proportion of good decisions under behavioral feedback (ρ=.390, p=.005), but the same was not the case under eye-feedback (ρ=.163, p=.258). In contrast, for buying decisions the analogous correlation was significant under eye-feedback (ρ=.516, p<.001), but not under behavioral feedback (ρ=.112, p=.440).

The way in which predicted goodness evolved throughout the study depending on the type of feedback is depicted in Fig. 7.
Figure 7. The evolution of predicted decision goodness throughout the study (starting with the trial in which feedback was introduced), separately for each type of decisions and depending on the type of feedback received (‘no feedback’ = Experiment 1). The piecewise step functions represent the average values of predicted goodness for each bin of 5 trials, while the straight lines represent the corresponding fitted linear trends.
3.2.3 Effect of Feedback on Decision Bias – Selling Decisions

We compared the average proportion of good selling decisions under eye-feedback, calculated separately for each subject, with one obtained in Experiment 1 and under behavioral feedback, with the latter comparison done separately for irreflective and reflective subjects (low vs. high CRT scores). We also carried out analogous between-group comparisons of the average PGR-PLR measures of the disposition effect (again, calculated separately for each subject). We used the Mann-Whitney test for all between-group comparisons. In addition, for the comparison of the two types of feedback, we conducted a mixed effects regression analysis of the subjects’ individual decisions.

We conducted the above analysis separately for buying and selling decisions, focusing on the latter and relegating the former to the Appendix. This was motivated by the fact that the influence of decision bias (disposition effect) on selling decisions is greater than that of the repurchase effect on buying decisions, as evidenced both by existing literature (Frydman et al., 2014; Frydman & Camerer, 2016) and the Experiment 1 results. Additionally, the work of Frydman and Rangel (2014) showed that selling decisions in particular may be debiased by reducing the visual saliency of the price level information. In line with this, we expected eye-feedback to be particularly useful in improving selling decisions, as in this case the bias is stronger and apparently associated with visual attention (see ‘Implications for Experiment 2’ in the Experiment 1 section). The remainder of the analysis is structured accordingly.
Between-group comparisons – proportion of good decisions. To begin with, the proportion of good selling decisions (of both irreflective and reflective subjects) under eye-feedback (M=53.5%) was higher than in Experiment 1 under no feedback (M=45.6%), U=1910.5, p=.019. In addition, the proportion of good selling decisions of irreflective subjects under eye-feedback (M=61.3%) was higher than the proportion of good selling decisions of irreflective subjects under behavioral feedback (M=51.6%), U=446.5, p=.048. Finally, the proportion of good selling decisions of reflective subjects under eye-feedback (M=45.0%) was lower than under behavioral feedback (M=63.9%), U=151, p=.005. Thus, the three comparisons are jointly significant based on the step-wise Bonferroni-Holm correction with successive p-values, in descending order, evaluated against respective alpha-thresholds of 0.05, 0.025 and 0.017.

Between-group comparisons – strength of the disposition effect. As we did for Experiment 1, we consider the difference between the proportion of gains realized (PGR) and the proportion of realized losses (PLR). The value of PGR-PLR (of both irreflective and reflective subjects) under eye tracking feedback (M=1.0%) was lower than in Experiment 1 under no feedback (M=26.5%), U=3276, p=.001. In addition, the value of PGR-PLR of irreflective subjects under eye-feedback (M=-17.2%) was lower than under behavioral feedback (M=10.2%), U=216.5, p=.025. Finally, the value of PGR-PLR of reflective subjects under eye-feedback (M=20.8%) was higher than under behavioral feedback (M=-22.3%), U=432.5, p=.003. Thus, the three comparisons are again jointly significant based on the Bonferroni-Holm correction.

Mixed-model estimation of the likelihood of selling. To compare the impact of the two types of feedback on individual selling decisions, we further estimated a mixed effects binary logistic regression model using data from Experiment 2, with random subject intercept and slope
effects and a binary dependent variable taking a value of one if the decision was to sell and zero otherwise. The reason why we used this variable instead of one indicating if the decision was good was that we wished to investigate how feedback type and the CRT score influence how subjects choose to trade in different circumstances, i.e. how they deal with stocks that are likely in the good state vs. those that are likely bad. For instance, we know from the earlier between-group comparisons that irreflective subjects are more likely to make good decision under eye-feedback than under behavioral feedback. But is this because eye-feedback makes them more likely to sell bad stocks, less likely to sell good ones, or both? Furthermore, a good decision to sell a stock that is almost certain to be in the bad state is not the same as one to sell a stock that is almost as likely to be in the good state as in the bad state. Hence, it is important to control for the circumstances in which decisions are made when evaluating them.

To this end, the independent variables in the regression model were: the probability $p_n$ of the stock being in the good state (‘prob-good’, rescaled to run from 0 to 1), a binary variable ‘high-CRT’ taking a value of one if the subject was classed as reflective and zero otherwise, and a binary variable ‘behav-feedb’ taking a value of one if the subject received behavioral feedback and zero if eye-feedback was given. We also included all interaction terms between those variables. The resulting fixed effect estimates are shown in Table 1.
Table 1. The fixed effect estimates of the mixed effects binary logistic regression, with the likelihood of selling an owned stock modeled as a function of the CRT result, feedback type and probability of the stock being in the good state.

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>Std. Err.</th>
<th>t</th>
<th>Sig.</th>
<th>Exp. Coeff</th>
<th>95% conf. int. (exp. coeff.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>intercept</td>
<td>0.511</td>
<td>0.329</td>
<td>1.55</td>
<td>.120</td>
<td>1.666</td>
<td>0.875</td>
</tr>
<tr>
<td>high-CRT</td>
<td>-1.807</td>
<td>0.481</td>
<td>-3.76</td>
<td>.000</td>
<td>0.164</td>
<td>0.064</td>
</tr>
<tr>
<td>behav-feedb</td>
<td>-0.938</td>
<td>0.466</td>
<td>-2.01</td>
<td>.044</td>
<td>0.391</td>
<td>0.157</td>
</tr>
<tr>
<td>prob-good</td>
<td>-1.696</td>
<td>0.630</td>
<td>-2.69</td>
<td>.007</td>
<td>0.183</td>
<td>0.053</td>
</tr>
<tr>
<td>high-CRT*behav-feedb</td>
<td>3.177</td>
<td>0.678</td>
<td>4.69</td>
<td>.000</td>
<td>23.97</td>
<td>6.347</td>
</tr>
<tr>
<td>high-CRT*prob-good</td>
<td>2.692</td>
<td>0.913</td>
<td>2.95</td>
<td>.003</td>
<td>14.76</td>
<td>2.467</td>
</tr>
<tr>
<td>behav-feedb*prob-good</td>
<td>1.947</td>
<td>0.886</td>
<td>2.19</td>
<td>.028</td>
<td>7.01</td>
<td>1.235</td>
</tr>
<tr>
<td>high-CRT<em>prob-good</em>behav-feedb</td>
<td>-5.936</td>
<td>1.292</td>
<td>-4.60</td>
<td>.000</td>
<td>.003</td>
<td>0.000</td>
</tr>
</tbody>
</table>
The regression estimates indicated that subjects we characterized as irreflective (those with low CRT scores) made better selling decisions under eye-feedback than under behavioral feedback. In particular, irreflective subjects ($\beta_{\text{high-CRT}=0}$) were less likely to sell bad stocks (those with prob-good close to zero) under behavioral feedback than under eye-feedback ($\beta_{\text{behav-feedb}=-0.938, \ p=.044}$). In Fig. 8, which illustrates the regression estimation results, this is reflected by the orange curve lying below the blue one for low values of the probability of the stock being in the good state (horizontal axis).

At the same time, when the probability that the stock is good was higher, irreflective subjects receiving eye-feedback ($\beta_{\text{high-CRT}=\beta_{\text{behav-feedb}=0}}$) were less inclined to sell it ($\beta_{\text{prob-good}=-1.696, \ p=.007}$), as reflected by the negative slope of the blue curve in Fig. 8. However, this change (slope) was significantly less negative when irreflective subjects received behavioral feedback ($\beta_{\text{behav-feedb*prob-good}=1.947, \ p=.028}$), which is represented by the fact that the slope of the orange curve is not as negative as that of the blue one, but rather almost completely flat. In other words, irreflective subjects who received eye-feedback, compared with those who received behavioral feedback, were more likely to sell bad stocks and more inclined to distinguish between bad and good stocks in favor of the latter, i.e. to choose to hold good stocks more often than bad ones. In this sense, irreflective subjects performed better under eye-feedback than under behavioral feedback, resulting in a higher proportion of good decisions observed in the earlier between-group comparisons.

However, the above gains from eye-feedback relative to behavioral feedback observed for irreflective subjects were significantly smaller, or indeed reversed, for the reflective ones ($\text{high-CRT}=1$). In particular, the (negative) effect of behavioral feedback on the likelihood of selling bad stocks observed for irreflective subjects was significantly more positive for reflective
ones ($\beta_{\text{high-CRT*behav-feedback}}=3.177$, $p<.001$). In Fig. 8, this was reflected by the fact that the intercept of the red curve was higher relative to that of the green one, in contrast with the intercept of the orange curve lying below the blue one. That is, unlike their irreflective counterparts, reflective subjects were more likely to sell bad stocks under behavioral feedback than under eye-feedback.

Similarly, the positive effect of behavioral feedback on the slope of the relationship with prob-good observed for irreflective subjects ($\beta_{\text{behav-feedback*prob-good}}=1.947$, $p=.028$) was reversed for their reflective counterparts ($\beta_{\text{high-CRT*prob-good*behav-feedback}}=-5.936$, $p<.001$). That is, the slope of the green curve was more positive relative to that of the red one than the slope of the blue curve was relative to the orange one. In other words, unlike their irreflective counterparts, reflective subjects were more inclined to favor good stocks over bad ones under behavioral feedback than under eye-feedback.

In fact, under eye-feedback ($\beta_{\text{behav-feedback}}=0$), reflective subjects performed worse than irreflective ones, in the sense that they were less likely to sell bad stocks ($\beta_{\text{high-CRT}}=-1.807$, $p<.001$) and less inclined to favor good stocks over bad ones ($\beta_{\text{high-CRT*prob-good}}=2.692$, $p=.003$). This was reflected by the intercept and slope of the green curve being lower and higher respectively relative to those of the blue curve.
Figure 8. The estimated relationship between the likelihood of selling (the dependent variable of the regression model in Table 1) and the probability of the stock being in the good state (continuous independent variable), for each of the four combinations of feedback and CRT classification (binary independent variables), with shaded 95% adjusted bootstrap percentile confidence intervals. Note that the probability of being in the good state, when not re-scaled to $[0;1]$, ranges from 0.2 to 0.8, because even a stock that is certain to have recently been in the good (or bad) state has a 20% chance of switching states between trials.
4 General Discussion

Comparing eye-feedback with no feedback. The results of Experiment 1 indicated that it is possible to distinguish between a good vs. bad decision-making process based on what information a person has looked at and in what order. Indeed, through a cross-validation procedure, we verified that this can be done out of sample, i.e. that a machine learning algorithm can accurately assess a subject’s decision process based on training examples of other subjects’ decisions.

The main finding of Experiment 2 was, in turn, that there were significant gains from using the algorithm trained on data from Experiment 1 to provide simple, good vs. bad feedback to a different group of subjects. To begin with, we found that subjects who received eye-feedback were indeed able to adjust their attentional patterns, as the predicted decision goodness was higher than in Experiment 1. In other words, they learned to visually examine information in a way that was associated with making good decisions in the first experiment. On the one hand, this might seem unexpected since, compared with conventional eye-movement modelling, subjects did not see a visualization of an ideal eye-movement pattern. On the other hand, they could infer it through trial and error and, unlike in conventional EMME, they were regularly informed whether or not their most recent eye-movements have been in line with the model.

Given the considerable difficulty of self-monitoring of eye-movements reported by existing research, the individual and frequent nature of eye-feedback could have played a part.

Crucially, the shift in attentional patterns induced by eye-feedback translated into improved decision-making: the proportion of good selling decisions was higher under eye-feedback than it was in Experiment 1 when no feedback was given. In addition, the disposition
effect was almost reduced to zero under eye-feedback, i.e. the tendency to realize gains more often than losses (measured by the PGR-PLR difference) vanished. Apart from learning to focus on the right pieces of information, this effect could have been driven by the fact that people are known to be able to infer others’ judgments from their observed eye-movements (Foulsham & Lock, 2015). Thus, knowing what good decision-makers looked at could have led subjects to infer the correct decision-making strategy, improving decisions via a less direct, theory-of-mind based route.

Interestingly, the above aggregate effects of eye-feedback mirrored what was observed in the behavioral feedback condition, where the predicted decision goodness also increased. Thus, using conventional means to teach people to make better decisions resulted in their eye-movement patterns becoming more similar to those that our algorithm associated with good decisions in the absence of feedback. Put simply, teaching people to make good decisions results in them learning how to make ‘good eye-movements’ as well, i.e. to visually process information like those other people who are ‘natural’ decision-makers, capable of deciding well even in the absence of feedback. Crucially, though, our results imply that the converse is also true, i.e. that it is possible to reverse this process by teaching people to make good eye-movements so as to induce them to make good decisions.

Comparing eye-feedback with behavioral feedback. At the same time, we found that the relative effect of the two types of feedback on decisions depends on subjects’ individual traits. On the one hand, reflective subjects performed better under behavioral feedback than under eye-feedback. This is not surprising when one considers the fact that the stocks’ price paths were governed by simple rules revealed to subjects together with the associated probabilities. This made behavioral feedback simple to understand relative to eye-feedback, setting the bar very
high for the latter. In fact, feedback that an individual investor would typically receive in the real-world might be closer to what is already provided by our ‘portfolio information screen’. That is, investors might reflect on how their capital gains evolve, but would not learn if their choices were Bayesian-optimal. By comparing with Experiment 1, we have seen that complementing such basic information with eye-feedback is beneficial to subjects.

On the other hand, the above considerations make it all the more interesting that the relationship between the two types of feedback was reversed in case of irreflective subjects, who were better off under eye-feedback than under behavioral feedback. These were the subjects with a particularly strong tendency for automatic, heuristic responses biased by irrelevant information. In the context of an experimental stock market setting, such a bias is likely to be accompanied by a tendency to ‘shut out’ relevant information that is nevertheless inconsistent with one’s prior choices (Kuhnen et al., 2017). We believe that this may have hindered the irreflective subjects’ learning from behavioral feedback since, in order to do so, they first had to suppress their pre-existing, incorrect way of dealing with the problem, sustained by biased attentional patterns. What is more, an irreflective subject driven by the disposition effect might focus on the price level information, selling when the capital gain is positive, and still receive good behavioral feedback in some of the trials, because even a stock that is almost certainly bad has a one in three chance of going up following its purchase, making selling it at a gain a good choice. In contrast, eye-feedback made it possible for irreflective subjects to learn to focus on the right pieces of information, particularly the price change history data, rather than the price level information that triggers their incorrect ‘gut’ response.

In essence, the above tradeoff between behavioral and eye-feedback can be explained via an analogy with trying to estimate a model of the relationship between a set of predictors (here,
the price change history and price level data) and a dependent variable (here, the Bayesian-optimal decision). For irreflective subjects, it is as if certain important predictors are omitted and irrelevant variables inserted in their place. Thus, even with several learning examples (provided by behavioral feedback), the estimated relationship is still biased. While eye-feedback can correct the misspecification of the predictor set, this comes at a cost of the dependent variable being subject to noise. This is because, in the absence of behavioral feedback, a subject can only learn if a decision was correct by observing the actual price changes that follow. As we have seen, there is an at least one in three chance that an ex ante good decision will prove incorrect ex post (because good stocks can fall etc.), making the observed price change an imperfect signal of decision goodness. Despite this, in case of irreflective subjects, the problems resulting from predictor misspecification are apparently so acute that a noisy dependent variable is a price worth paying.

Apart from the above results, confirming our hypotheses, we also obtained one that, while not strictly contradicting those hypotheses, was rather surprising. Specifically, under eye-feedback, reflective subjects performed worse than irreflective ones. One potential explanation of this, based on reflective subjects’ comments during debriefing sessions, is that they may have perceived themselves as equipped with above-average decision skills, and so did not respect their peers as a source of feedback (see Narciss & Dünnebier, 2010 for a discussion of the related issues, and note that subjects were aware that behavioral and eye feedback were based on an objective probabilistic standard and the subjective performance of others respectively). Success in tasks involving learning from peers opportunities has already been shown to be positively related to the propensity to imitate others (Wisdom, Song, & Goldstone, 2013), while ignoring the instruction and consistently getting bad feedback could have caused distraction and
performance stress. This may be related to existing research on the ‘illusion of competence’ and ‘curse of expertise’ (Fisher & Keil, 2016).

Additionally, an observation that prompted our use of the CRT test in Experiment 2 was that a significant proportion of Experiment 1 subjects performed well despite making eye movements generally associated with bias. If, as we hypothesized, those subjects were characterized by relatively high cognitive reflection, then eye-feedback would unnecessarily prompt them to switch from a way of collecting information that, for them, works very well, to one that might be less suited to their cognitive traits.

The above observations could partly explain the lack of significant correlation between predicted decision goodness and the proportion of good selling decisions under eye-feedback. This lack of correlation was contrary to the relationship between predicted and actual goodness established in Experiment 1. One could interpret this discrepancy through an analogy with a recent work showing that visual attention to computational tools is similar in experts vs. novices, because the design of these tools induces the ‘right’ eye-movement patterns irrespective of expertise (Srinivasan, Wagner, C. Frank, & Barner, 2018). Similarly, the presence of eye-feedback creates a ‘level playing field’, by changing the decision environment’s architecture so that even subjects with little natural capacity for good decisions learn to follow the same attentional patterns as their more capable peers.

**Potential Applications.** Although we evaluated the proposed technique in the context of a specific decision problem, we believe that the proof of concept which we provide could be interesting for a number of reasons. In particular, technologies are already in place making it possible to gather eye-data via in-built cameras of personal devices like laptops or smartphones (Krafka et al., 2016; Semmelmann & Weigelt, 2017), or to approximate it through even more
basic techniques (Reisen, Hoffrage, & Mast, 2008), and then pool the collected data using Internet services such as Amazon's Mechanical Turk (Xu, Ehinger, Zhang, & Finkelstein, 2015). Importantly, our model of ideal eye-movements is automatically crowdsourced rather than provided a priori by the instructor with experts’ help. This means that the debiasing effect we obtained relies on ‘learning from peers’ (Bursztyn, Ederer, Ferman, & Yuchtman, 2014; Foucault & Fresard, 2014), instead of costly expert guidance. Thus, a potential future implementation of the proposed technique in a real-world setting is likely to be comparatively cheap, despite allowing for the provision of frequent, individual process feedback.

In this context, one should also note that, while the machine-learning algorithm we used to provide eye-feedback was relatively complex, a less sophisticated one, based on fixation times, was not much worse in terms of distinguishing between good vs. bad decisions (as illustrated in Figure 3). The aim of the current project was to show that some form of eye-feedback might benefit the decision-makers, and we opted for its most effective variant as the surest way to demonstrate this. However, future research might show that even considerably less complex techniques, e.g. ones based on dwell times alone, could be similarly effective, and so more cost efficient, because they could work well with limited data provided by cheap, home devices.

In fact, the demonstrated potential of eye-feedback could shed light on several important, more general questions of significant practical relevance. First, latest research shows that even highly complex tasks can be efficiently solved via ‘crowdsourced science’, i.e. combining inputs from a large number of amateur human agents, each using simple and intuitive decision heuristics (Sørensen et al., 2016). Our work suggests that it might be possible to crowdsource not
only answers to specific questions, but also the way in which a problem of a certain structure should be solved, i.e. the recipe rather than the final product.

Second, for all the well-known benefits of peer-to-peer interaction in learning, creating instructional software that would facilitate or simulate such interactions is challenging (Howard, Di Eugenio, Jordan, & Katz, 2017). Our findings indicate that it might be fruitful to allow for the exchange between peers of the attentional ‘recipes’ they use to solve problems, rather than information that they deliberately choose to communicate. In this way, good decision-making could potentially be reverse-engineered by studying attention. That is, if we extract the way in which good decision-makers visually process information, and then teach bad decision-makers, through feedback, to process it in the same way, then the latter might be able to self-regulate their attention accordingly, leading to a performance improvement.

Finally, latest research demonstrates how elements of human expertise, encoded in eye-tracking data, can be built into a computational model to support complex group decisions (Busey, Nikolov, Yu, Emerick, & Vanderkolk, 2017). Based on our results, we believe that decision support tools based on a blend of artificial intelligence and human cognitive process tracing might prove similarly effective in reducing individual decision bias.

Scope and Limitations. Naturally, in a real-world context, eye-feedback might be harder to implement than in the present laboratory setting. In particular, it may then not be possible to formulate an objective criterion of a decision being good ex-ante, i.e. making good use of the information available at the time, akin to our ‘Bayesian optimal’ criterion (this would apply also to any laboratory investment task using historical data instead of a simulated price process, as in, e.g. Lejarraga et al., 2016). To address this problem, one could use approximate ex-post measures of decision goodness, based on subsequent observed price changes. A cluster of
scanpath representations would then be associated with good decision-making if, on average, the associated decisions proved correct ex-post, leading to profitable outcomes. The resulting increase in noise and reduction in eye-feedback quality could at least partly be offset with a sufficiently high amount of data, and should, in any case, be less detrimental for eye-feedback than for behavioral feedback. Specifically, from the decision-maker’s perspective, the experience of eye-feedback would not change much, because she would still be taught to view information in the way the machine learning algorithm considers best (the underlying increase in noise being handled by the algorithm ‘in the background’). In contrast, in case of behavioral feedback, the person would have to deal with the noise herself, and learn from feedback that is stochastic and often inconsistent across repeated choices.

In the same vein, the decision screens we used were stylized and the same for all subjects, which would not be the case in reality. Nevertheless, our eye-tracking area of interest specification was extremely simple, while recent advances in computer vision enable real-time tracking of AOI’s even when their position on the screen is not fixed (see e.g. Friedrich, Rußwinkel, & Möhlenbrink, 2017). In other words, a machine learning algorithm could work out what a person is looking at before deciding if she does it ‘in the right way’.

Relatedly, future research might explore using feedback based not only on the order in which the AOIs were visited, but also on other information, like the duration of each visit. In particular, we found that, compared with an algorithm based on proportional attention to the key AOIs, the temporal difference learning technique allowed for an assessment of decision goodness that was more accurate, but still quite prone to errors (recall Figure 5). Perhaps a method incorporating both the ordering and duration of AOI visits, such as the ScanMatch technique (Cristino, Mathôt, Theeuwes, & Gilchrist, 2010) could improve feedback performance.
A separate issue is that our sample of subjects, drawn from a student population, is relatively uniform. At the same time, our method thrives on heterogeneity, as it identifies individuals who are particularly good at making decisions and teaches their less capable peers to imitate their visual processing patterns. Thus, we would argue that the uniform nature of the subject pool actually increases the strength of the current demonstration, as it suggests that, even in a world devoid of experts, one could still identify individuals with enough natural ability for their less capable peers to benefit from imitating their information acquisition patterns. At the same time, given the relatively small number of Experiment 2 subjects per cell, we acknowledge the need to replicate our findings, ideally using an increased sample size.

Lastly, a recurring issue in domains like financial education is that behavioral changes it brings about tend to be short-lived after the intervention is withdrawn. On the one hand, the time span that we consider here is short and it is difficult to gauge the long-term impact of our method. On the other hand, its potentially low-cost nature (compared with support offered by professional educators) means that feedback could potentially be provided indefinitely, thereby avoiding the problem of its post-withdrawal effectiveness.

5 Conclusions

We demonstrated a new way of using eye movement modeling in improving decisions, one that is both effective, in that it provides frequent individual feedback at the point of decision-making, and cheap, as it relies on peer-to-peer learning instead of costly inputs of professional educators. In contrast with conventional eye movement modelling, the model of ideal gaze patterns is crowdsourced rather than provided by experts, while the decision-makers are repeatedly informed if their eye-movements are consistent with the model.
Naturally, being the first example of using feedback based on process-tracing techniques and machine learning, the present work should be seen as an early proof of concept, demonstrating the potential of the technique within a specific decision setting. Nevertheless, we think that it is interesting to see that people can learn, based on simple, binary feedback, to emulate others’ eye-movements, and that emulating the eye-movements that usually accompany good choices can, in turn, improve one’s decisions. In other words, one can reverse engineer a good decision process from the accompanying attentional patterns.

Given the mentioned proof of concept nature of this work, we hope that it would provide grounds for further research. In particular, given the demonstrated tradeoff between process and outcome feedback (moderated by individual cognitive traits), it might be fruitful to experimentally evaluate the effectiveness of a combination of the two types of feedback: with eye-feedback teaching people what information to focus on, and behavioral feedback helping them understand how that information determines the correct choice.

More generally, we believe that our results could potentially inspire new solutions in the domain of crowdsourced science. They suggest that, by combining inputs from a large pool of amateurs, it may be possible to crowdsource not only answers to specific, complex questions, but also the way in which a problem of a certain structure should be solved, thereby crowdsourcing the recipe rather than the final product. What is more, the obtained recipe can apparently be ‘transferred’ to other people via a decision support tool based on a blend of artificial intelligence and cognitive process tracing. We hope that, given the crucial role of attention in human choices, further research will demonstrate the effectiveness of similar debiasing techniques in other domains of decision-making.
References


Introduction


Appendix: Effect of Feedback on Decision Bias – Buying Decisions

The proportion of good buying decisions (of both irreflective and reflective subjects) under eye tracking feedback (M=59.4%) was higher than in Experiment 1 under no feedback (M=51.6%), U=1939, p=.031. Additionally, the proportion of good decisions of irreflective subjects under eye-feedback (M=65.6%) was higher than under behavioral feedback (M=56.0%), U=429, p=.098. Finally, the proportion of good decisions of reflective subjects under eye feedback (M=52.8%) was lower than under behavioral feedback (M=67.1%), U=174,5, p=.019. However, based on the step-wise Bonferroni-Holm correction, none of the three comparisons is significant, although we see a trend towards significance.

The value of PDR-PUR (of both irreflective and reflective subjects) under eye tracking feedback (M=-17.6%) was lower than in Experiment 1 under no feedback (M=5.8%), U=1740, p=.005. In addition, the value of PDR-PUR of irreflective subjects under eye-feedback (M=-31.4%) was lower than under behavioral feedback (M=-13.9%), U=410, p=.191. Finally, the value of PDR-PUR of reflective subjects under eye-feedback (M=-2.6%) was higher than under behavioral feedback (M=-31.0%), U=180,5, p=.026. Thus, only the first of the three comparisons is significant based on the step-wise Bonferroni-Holm correction, although for the last comparison we see a trend towards significance.

As for selling decisions, we estimated a mixed effects binary logistic regression and a binary dependent variable taking a value of one if the decision was to buy and zero otherwise. We included the same fixed and random effects as for selling decisions, and the resulting estimates are shown in Table 2.
Table 2. The fixed effect estimates of the mixed effects binary logistic regression, with the likelihood of buying a stock modeled as a function of the CRT result, feedback type and probability of the stock being in the good state.

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>Std. Err.</th>
<th>t</th>
<th>Sig.</th>
<th>Exp. Coef</th>
<th>95% conf. int. (exp. coeff.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-1.727</td>
<td>0.358</td>
<td>-4.82</td>
<td>.000</td>
<td>0.178</td>
<td>0.088 - 0.359</td>
</tr>
<tr>
<td>high-CRT</td>
<td>0.525</td>
<td>0.514</td>
<td>1.02</td>
<td>.307</td>
<td>1.690</td>
<td>0.617 - 4.629</td>
</tr>
<tr>
<td>behav-feedb</td>
<td>1.038</td>
<td>0.505</td>
<td>2.05</td>
<td>.040</td>
<td>2.824</td>
<td>1.047 - 7.615</td>
</tr>
<tr>
<td>prob-good</td>
<td>3.200</td>
<td>0.667</td>
<td>4.80</td>
<td>.000</td>
<td>24.53</td>
<td>6.636 - 90.67</td>
</tr>
<tr>
<td>high-CRT*behav-feedb</td>
<td>-1.257</td>
<td>0.728</td>
<td>-1.73</td>
<td>.084</td>
<td>0.284</td>
<td>0.068 - 1.185</td>
</tr>
<tr>
<td>high-CRT*prob-good</td>
<td>-2.349</td>
<td>0.956</td>
<td>-2.46</td>
<td>.014</td>
<td>0.095</td>
<td>0.015 - 0.622</td>
</tr>
<tr>
<td>behav-feedb*prob-good</td>
<td>-1.965</td>
<td>0.933</td>
<td>-2.11</td>
<td>.035</td>
<td>0.140</td>
<td>0.023 - 0.872</td>
</tr>
<tr>
<td>high-CRT<em>prob-good</em>behav-feedb</td>
<td>4.349</td>
<td>1.356</td>
<td>3.21</td>
<td>.001</td>
<td>77.43</td>
<td>5.419 - 1106.2</td>
</tr>
</tbody>
</table>

Figure 9. The estimated relationship between buying likelihood and the probability of the stock being in the good state (without rescaling), with shaded 95% confidence intervals.