A pupillary index of susceptibility to decision biases

Eran Eldar[a,b,*], Valkyrie Felso[c,d], Jonathan D. Cohen[c,d], Yael Niv[c,d]

[a] Psychology and Cognitive Sciences Departments, The Hebrew University of Jerusalem
[b] Max Planck University College London Centre for Computational Psychiatry and Ageing Research
[c] Princeton Neuroscience Institute, Princeton University
[d] Psychology Department, Princeton University
[*] Correspondence should be addressed to: Department of Psychology
The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 9190501, Israel
eran.eldar@mail.huji.ac.il

Abstract

The demonstration that human decision making can systematically violate the laws of rationality has had wide ranging impact on the fields of economics and psychology. However, the cognitive processes that give rise to irrational biases are still poorly understood. In this study, we use a pupillary index to arbitrate between two predominant existing hypotheses – the hypothesis that biases result from fast effortless processing and the hypothesis that biases result from more extensive integration. While effortless processing is associated with smaller pupillary responses, more extensive integration has been shown to be associated with larger pupillary responses. Thus, we test the relationship between pupil responses and choice behavior on six different foundational decision-making tasks classically used to demonstrate irrational biases.

Introduction

In certain well-described scenarios, human decision making exhibits systematic deviations from rational behavior. For instance, exactly how a problem is described can determine whether a particular option is more or less likely to be chosen, even when equivalent information is provided by the different descriptions (e.g., "framing effect"[1]). The discovery and characterization of such biases has had substantial impact on the fields of psychology and behavioral economics[2]. However, the mechanisms underlying biased decision making remain widely debated.

The dominant paradigm posits that biased decisions arise from a fast and effortless intuitive process, which can be corrected via slower, effortful, deliberation[2,3]. However, a separate line of work proposes essentially the opposite – that biases arise from a gradual process of evidence integration[4-11]. While these two theories are not necessarily mutually exclusive, each theory provides a different account for why some people may be more biased than others. Specifically, the former theory suggests that biased decision makers employ an effortless process, whereas the latter theory suggests they employ more extensive integration (see **Supplementary Material** for an example of a computational model illustrating the latter mechanism).

Critically, these two explanatory factors, low effort and extensive integration, are known to be associated with opposite changes in pupil diameter. It is well established that lower effort is accompanied by lower pupillary responses[12]. On the other hand, recent studies show that people with higher pupillary responses integrate more extensively different aspects of available information[13–15]. This latter finding is among a set of neural and behavioral results explained by an hypothesized relationship between high pupillary responses, lower levels of sustained locus coeruleus-norepinephrine function, and low neural gain[13,16-20]. In previous theoretical work, we simulated low levels of gain (which means that incoming neural signals have a weaker impact on the postsynaptic neuron) and showed that the result of this parameterization is a more prolonged integration of information for decision making, which allows a broader set of sources of information to influence the decision, including sources that are less salient or of secondary importance[14]. Such inclusive integration may be necessary to allow weak biasing influences, which are typically marginal or even irrelevant to the problem at hand, to exert their effect.

Thus, analyzing decision makers' pupil diameter could tell us which mechanism—an automatic effortless process or extensive integration—is likely responsible for generating biased decisions. Further, understanding the relationship between individual differences in susceptibility to decision biases and pupil dynamics can provide a simple, non-invasive method for measuring an individual's tendency to be biased by the way a problem is described.

Here we test human participants on six well-established decision-making tasks from the heuristics and biases literature, while measuring their pupil dilation responses during performance of the tasks. If neither of the theories outlined above is correct (or if biases on different tasks are generated by different mechanisms), we should not see any overall relationship between pupil response and biases. However, if one these theories consistently explains individual differences in biased decision making, pupil response measurements should distinguish between participants who are more susceptible to biased decision making and those who are relatively immune to these manipulations. A negative relationship between pupil response and biases would support the long-standing belief that biases are generated by an effortless automatic decision process, whereas a positive relationship would indicate that biases are produced by gradual integration of evidence. Equally important, the latter result would suggest a potential role for low levels neural gain in facilitating the manifestation of decision biases. The only results of this experiment that would be less than illuminating are a mix of relationships between pupillometry and susceptibility to biases across tasks. To validate our pupillometric measurements and to measure an additional complementary index of neural gain we include one minute of a classic oddball task between every two test tasks. The reliable dilation of the pupil in response to oddballs[19,20] will serve as a positive control. Further, response times on such perceptual discrimination tasks can be expected to reflect neural gain, as indicated by computational modeling and experimental evidence[14]. Thus, a neural gain account of decision biases would be further supported by the association of biased decisions with slower responses to oddballs.

Methods

Participants. 120 participants will be recruited from the greater Princeton area. The sample size was determined via a bootstrapping-based power analysis of pilot data (see below). Inclusion criteria are age 18 to 35 and compatibility with pupillometry, as evidenced by successful calibration of the eye tracker. Participants will give written informed consent before taking part in the study, which is approved by the university's institutional review board. Participants will receive either course credit or compensation of $12 per hour for participation.

Power analysis. To determine the sample size, we used data from 44 pilot participants to compute the expected probability of meeting the weak and strong criteria in support of the study's hypotheses (detailed under **Statistical analysis**) for different numbers of participants. Expected probabilities were computed by performing the analysis on 1000 datasets, each of which constructed by sampling participants with replacement from the pilot data. The power analysis showed that a sample size of 120 participants provides a 95% probability of finding strong support for the study's hypothesis, given the effect size found in the pilot data (**Figure 1**). While smaller effect sizes might be of theoretical importance, an effect size commensurate with that found in the pilot data would be necessary for pupillary measurement to reliably predict susceptibility to decision-making biases.

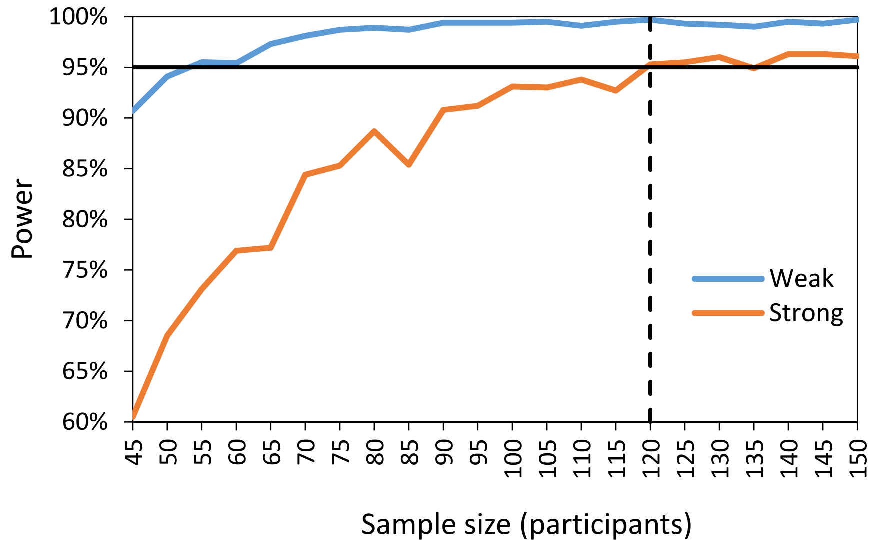

Figure 1. Power analysis. Expected probability of meeting the weak and strong criteria in favor of the study's hypotheses for different numbers of participants (see **Statistical analysis** for a specification of the criteria). Probabilities were computed by analyzing 1000 datasets, each of which constructed by sampling participants with replacement from the pilot data. Horizontal line: 95% power. Vertical line: minimal sample size required to achieve 95% power.

Stimuli. Stimuli were generated using the Processing programming environment[21]. To minimize luminance-related changes in pupil diameter, we will first identify colors that are isoluminant with the background by having participants perform the flicker-fusion procedure[22] on the display system that will be used in the experiment. The colors of the experimental stimuli will then be automatically adjusted accordingly, to achieve subjective isoluminance in the conditions of the testing room, for each participant. Stimuli will be presented on a computer screen using MATLAB software (MathWorks) and the Psychophysics Toolbox[23].

Experimental design. Each participant will perform six experimental tasks, each aimed at inducing a different bias. To facilitate comparisons between participants, all participants will perform all tasks in the order in which the tasks are described below. The experiment will last approximately 1 hour. Unless otherwise noted, questions will appear on the screen until the participant enters their answer using a keyboard (i.e., there will be no time restrictions for providing an answer). To allow sufficient time for pupillary responses to be resolved, questions will be separated by random inter-trial intervals, 7 to 9 s long (uniformly distributed), during which only a fixation cross will appear on the screen.

Task 1: Anchoring task[24]. Participants will answer two questions about each of 7 quantities (e.g., the height of the Eiffel tower). They will first be asked to indicate whether the quantity is greater ('1' keyboard key) or smaller ('2' keyboard key) than an anchor value. Once the participant responds, the first question will disappear from the screen, and the participant will immediately be asked to estimate the quantity by typing it using the keyboard and then pressing ENTER. Each quantity will be coupled with a low anchor for half of the participants and with a high anchor for the other half. Each participant will be presented with a low anchor for half (3 or 4) of the quantities, and with a high anchor for the other half. Quantities and calibrated anchor values are taken from a previous study[25], including: length of the Mississippi river, population of Chicago, number of babies born per day in the US, height of mount Everest, pounds of meat eaten by an American per day, year the telephone was invented, and maximum speed of a house cat. Participants' estimates will be normalized to a common scale (0 = lowest estimate, 1 = highest estimate) by subtracting the lowest estimate and then dividing by the highest resulting estimate. The group mean estimate, averaged over both types of anchors, provides a measure of what an average person who is not affected by the anchors is likely to answer. The anchoring effect will therefore be quantified by the deviation of an estimate in the direction of the anchor relative to the mean estimate provided by the whole study sample. Estimates whose distance from all other participants' mean estimate is more than ten times the range of the other participants' estimates will be excluded as outliers.

Task 2: Persistence of Belief task[26]. Participants will be presented with two urns, each filled with 10 colored balls (**Figure 2a**). One urn will contain 3 red balls, 2 green balls, 2 blue balls, 2 brown balls and 1 purple ball, and the other urn will contain 2 red balls, 3 green balls, 1 blue ball, 2 brown balls and 2 purple balls. Participants will then be shown a sequence of 90 balls, which they will be told were sampled with replacement from one of the urns. Each sampled ball will fall from the top of the screen, horizontally centered, until it settles near the bottom of the screen, and it will then disappear. Balls will follow one another in sequence without a break (3.3 s per ball), while the two urns are presented on the left and right sides of the screen. Every 5 samples (balls), participants will be asked to indicate using an appropriately-labeled horizontal sliding bar which urn they think the sequence was sampled from. Participants will be instructed to indicate their degree of certainty by means of the precise position of the bar, where a center position corresponds to total uncertainty. Each question will be followed by an inter-trial interval. The sequence of balls will be set up so that the first 30 balls favor one of the urns as their source with a probability of 0.95, and the next 60 balls favor the other urn to a similar degree (per 30 balls). Therefore, it is optimal to favor one urn after 30 balls, be indifferent after 60 balls, and favor the second urn after 90 balls (**Figure 2b**). Accordingly, an optimal observer would be indifferent on average during the last 60 balls. However, the biasing impact of an initially formed belief on the interpretation of later evidence, akin to a framing effect, is expected to slow down belief reversal. Thus, a persistence-of-belief effect will therefore be quantified by the degree to which each participant's average response during the last 60 balls favors the initially-favored urn. The initially-favored urn will be counterbalanced across participants. Data from participants who do not favor the correct urn during the first 30 balls will be excluded from analysis.

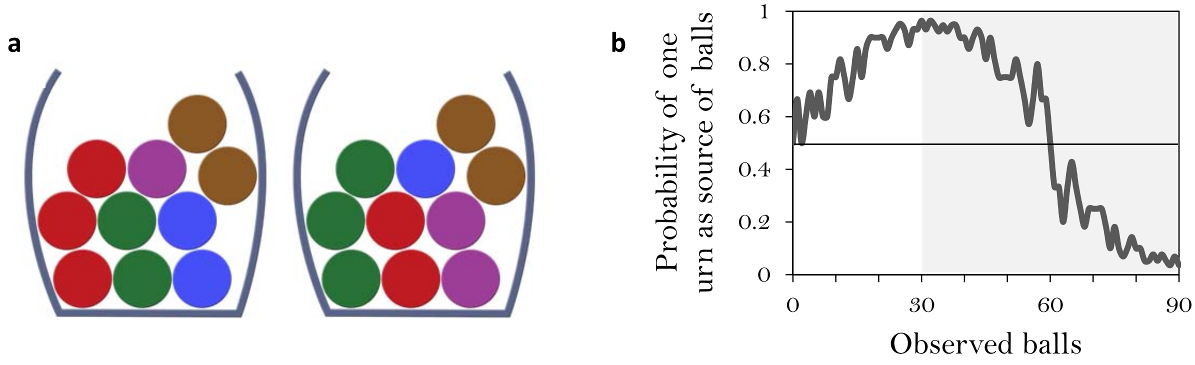

Figure 2. Persistence of Belief task. (**a**) The two urns presented to participants contain different proportions of balls of different colors. Balls drawn are pre-determined such that at first it seems that they are drawn from one urn, whereas later evidence suggests the other urn. (**b**) Probability of one urn being the source of the sequence of balls as the sequence progresses, determined by the relative likelihood of each of the balls coming out of the urn, given the contents of both urns. On average, between trials 30-90 this specific sequence is equally likely to come from either of the urns (starting from 95% likely to come from one urn, and symmetrically changing to 95% likely to come from the other urn). Note that even if participants did not reach 95% certainty in the first 30 trials, as long as their updates are symmetric, they should go back down to 0 around trial 60 and to the opposite asymptote at about trial 90, meaning that on average there should be indifference on trials 30-90.

Task 3: Attribute framing task[27]. Participants will be asked to rate ground beef products, gambles, and students' performance, whose attributes are framed either positively or negatively. In the ground beef task, participants will be asked to imagine that they are having a friend over for dinner and they are about to make their favorite lasagna dish with ground beef. They will then be asked to rate how satisfied they would be purchasing each of 4 ground beef products, described in terms of price per pound ($2.7 and $3.3), and either percentage lean (80% and 90%, positive frame) or percentage fat (20% and 10%, negative frame). In the gambles task, participants will be asked to imagine that they have $10 and can either keep the $10 or pay the $10 to take a gamble. They will then be asked to rate how likely they are to take each of 3 gambles, described in terms of amount to be won ($50, $100 and $200) and either probability of wining (20%, 10% and 5%, positive frame) or probability of losing (80%, 90% and 95%, negative frame). In the student performance task, participants will be asked to evaluate each of 2 students on the basis of midterm exam and final exam performance, described in terms of either percent correct (50% and 70%, positive frame) or percent incorrect (50% and 30%, negative frame). The attributes of an item will remain on the screen until the participant finishes rating the item by adjusting an appropriately labeled vertical sliding bar and then pressing ENTER. Each item will be framed positively for half of the participants, and negatively for the other half. For a given participant, all items of a particular type will be similarly framed (i.e., either positively or negatively), so as to minimize awareness of the framing manipulation, but framing will be varied within participants across item types. As in the Anchoring Task, the framing effect will be quantified for each item type by the deviation of a participant's mean rating from the from the overall mean rating, in the direction of the frame (i.e., upwards for positive frames, and downwards for negative frames).

Task 4: Risky Choice framing task[28]. Participants will face two different scenarios, a medical scenario and a fire scenario, and they will be asked to indicate using a sliding bar which of two available actions they would choose in each scenario. One action will have a certain outcome and the other an uncertain outcome, both of which will be framed in terms of either gains or losses (counterbalanced across participants). Scenarios will be described in full as done previously[28]. In the medical scenario, which concerns the treatment of a deadly disease at an island inhabited with 600 inhabitants, participants will be asked to choose between gain-framed outcomes '300 people will be saved' and 'a 50% chance that 600 people will be saved and a 50% chance that none of the people will be saved', or between loss-framed outcomes '300 people will die' and 'a 50% chance that 600 people will die and a 50% chance that none of the people will die'. In the fire scenario, which concerns the treatment of fires threatening 9000 acres of forest, participants will be asked to choose between gain-framed outcomes '3000 acres of forest will be saved' and 'a 60% chance that 5000 acres will be saved and a 40% chance that no forest under threat will be saved', or between loss-framed outcomes '6000 acres of forest will be lost' and 'a 60% chance that 4000 acres will be lost and a 40% chance that 9000 acres will be lost'. For each question, the attributes of the first option (as described above) will appear on the left side of the screen, and the attributes of the second option will appear on the right side of the screen. These details will remain on the screen until the participant indicates their preference by adjusting an appropriately labeled horizontal sliding bar and then presses ENTER. As for the Anchoring and Attribute framing tasks, the framing effect will be quantified as the deviation of a participant's preferences from the overall mean rating, in the direction of the frame (i.e., towards the certain outcome in the gain frame and towards the uncertain option in the loss frame, in line with people's well-documented risk-aversion in the gain domain and risk-seeking in the loss domain[29]).

Task 5: "Task Framing" task[30]. Participants will face 5 different problems, concerning various subjects such as child custody, vacation choice, ice-cream choice and gambling. Each problem will involve one option that has more positive and negative attributes (the 'enriched' option) and one option that has fewer positive and negative attributes (the 'impoverished' option). In each problem, half of the participants will be asked to choose one of the two options, and the other half will be asked to reject one of the two options. For example, in one problem participants will be asked to imagine that they serve on the jury of an only-child sole-custody case following a relatively messy divorce, and they have to make a decision based entirely on the following few observations. Parent A: average income, average health, average working hours, reasonable rapport with the child, relatively stable social life (this parent has no particularly positive or negative attributes). Parent B: above-average income, very close relationship with the child, extremely active social life, lots of work-related travel, minor health problems (this parent has 3 positive and 2 negative attributes). Half of the participants will be asked to which parent they would award sole custody of the child, while the other half will be asked which parent they would deny sole custody of the child. Full description of the other problems can be found elsewhere[30] (problems 1, 2, 4, 5 and 6). Participants will be asked to report their preferences in the same way as in the Risky Choice framing task above (that is, by adjusting a horizontal slider bar with the two options displayed on each side of the bar). The task frame (award vs. reject) will be varied within participants across questions. The task-framing bias manifests in people's tendency to choose the enriched option as opposed to the option they have less information about. Because the enriched option has more positive and more negative attributes, the bias manifests similarly regardless of whether participants are asked to express a preference for one option (i.e., award frame) or rejection one option (i.e., reject frame). Thus, the framing effect will be quantified by the degree to which each participant chooses the enriched option (i.e., Parent B) more frequently than the impoverished option (i.e., Parent A).

Task 6: Sample-Size Neglect task[31]. Participants will be asked to imagine that they are tossing a biased coin and recording how often the coin lands heads and how often the coin lands tails. They know that the coin is bent and tends to land on one side 3 out of 5 times, but they do not know if this bias is in favor of heads or in favor of tails. Participants will then be presented with 10 different sets of results (number of heads and number of tails), in which the heads always outnumbered the tails, and they will be asked to indicate using a vertical sliding bar how *certain* they are given each set that the coin is biased in favor of heads. The top end of the bar will be labeled with "completely certain that coin favors heads", and the bottom end with "completely uncertain that coin favors heads". Each set of results will remain on the screen until the participant finishes adjusting the bar and presses ENTER. Sets of results will be similar to those used previously[31].

As shown by Griffin & Tversky[31], the probability that the coin is biased in favor of heads according to Bayes' rule is:

$$p(H|D) = e^{(h-t)\log\frac{3}{2}} \qquad (1)$$

where h is the number of heads and t is the number of tails. This expression is equivalent to

$$p(H|D) = e^{n\frac{(h-t)}{n}\log\frac{3}{2}} = e^{n\frac{(h-t)}{(h+t)}\log\frac{3}{2}} \qquad (2)$$

which depends on the sample size (i.e., the number of outcomes, n) and on the observed ratio of heads and tails ($\frac{h-t}{h+t}$). Previous work has shown that people tend to overweigh the ratio component at the expense of the sample size component (sample-size neglect[31]). Thus, to measure this bias for an individual participant, we will regress the participant's estimates against the true probabilities (Eq. 1) as well as against the ratio component ($e^{\frac{h-t}{h+t}\log\frac{3}{2}}$), and compare the two resulting regression coefficients (β_{true} and β_{ratio}). All inputs to the regression analyses will be z scored so as to produce normalized coefficients, such that perfect correlation between the participant's ratings and the true probability would yield $\beta_{\text{true}} = 1$ and $\beta_{\text{ratio}} = 0$, while complete reliance on the ratio between heads and tails would yield $\beta_{\text{true}} = 0$ and $\beta_{\text{ratio}} = 1$. Thus, the sample-size neglect will be computed for each participant as $1 - \beta_{\text{true}} + \beta_{\text{ratio}}$.

Data from participants for whom β_{true} and β_{ratio} are lower than 0, or who report higher certainty given 3 heads and 2 tails, than given 7 heads and 2 tails, will be excluded from the analysis. The former criterion would indicate the participant did not give reasonable answers, and the latter criterion would suggest specifically that the participant mistakenly looked for a ratio that best matches 3 to 2.

Oddball task. To assess reaction time and pupillary responses in a uniform manner throughout the experiment, and as a positive control to our other findings, we will use a shortened version of an auditory oddball task, in which robust anti-correlations between pupil response and baseline pupil diameter have previously been demonstrated[19,20]. Participants will be presented with a sequence of 60-ms sinusoidal tones, of two possible frequencies: 1000 Hz tones, which will be designated as the target, and 500 Hz tones, which will be designated as non-targets. Participants will be told to respond with a keypress only when the target tone is sounded. Inter-tone intervals will be drawn uniformly between 2.1 to 2.9 seconds. To allow the pupil diameter to return to baseline, the stimuli will be ordered such that target tones will always be spaced between at least three non-target tones on each side. Target tones will make up 20% of the tones. Results of pupil diameter response to the oddball items will be analyzed to verify reliable pupillometry measurements. As in previous studies[19], we will exclude from analysis trials in which a participant responded to a non-target tone (false positive), did not respond to a target tone (miss) or responded within 100 ms of target presentation (quick response).

Participants will perform a total of seven oddball task blocks, such that oddball blocks alternate with the six decision making tasks. Each block will consist of 25 tones (5 of them oddballs). Oddball reaction time and pupillary response will be computed for each decision-making task based on the oddball blocks that immediately precede and follow the task (that is, based on a total of 50 tones / 10 oddballs). These measures will be used for complementary analyses identical to the main analyses described below, but replacing the task pupillary responses with the oddball reaction times and pupillary responses.

Eye tracking. A desk-mounted SMI RED 120Hz eye-tracker (SensoMotoric Instruments Inc., MA) will be used to measure participants' left and right pupil diameters at a rate of 60 samples per second while they are performing the behavioral tasks with their head fixed on a chinrest. At the beginning of the experiment, a baseline measurement of pupil diameter at rest will be taken for a period of 45 s. Pupil-diameter data will be analyzed in MATLAB as in previous work[13,14]. First, the data will be processed to detect and remove blinks and other artifacts. For this purpose, artifactual diameter samples will be identified as those lower than 66% of, or higher than 150% of, the median non-zero sample, as well as those samples that differ from adjacent samples by more than 10%. Samples recorded between 33 ms before to 100 ms after an artifact will also be designated as artifactual. All artifactual samples will be replaced by linear interpolation. For each task and each question, baseline pupil diameter will be computed as the average diameter over a period of 1 s prior to presentation of the question. Based on an examination of the pilot data (**Supplementary Figure 2a**), we determined that in the six decision-making tasks, pupil-dilation response will be computed as the peak diameter recorded during the period between 1s and 6s following presentation of the question, minus the preceding baseline diameter. For the oddball task, pupil responses are shorter (**Supplementary Figure 2b**), and thus, the peak diameter will be assessed between 0.4s and 2s following stimulus onset. All pupil-dilation responses will be normalized by the pre-experiment baseline pupil diameter. Questions and oddball trials for which more than half of the pupil measurements are affected by artifacts will be considered invalid and excluded from the analysis. Participants with fewer than two valid (i.e., mostly artifact free) questions in a given task will be excluded from the analysis of that task.

Statistical analysis. For each task, we will divide participants into tertiles of low, medium and high mean pupil dilation. This will allow us to visualize the degree to which each group exhibited a significant bias on each task. Then, to test for an overall relationship between pupil response and biases across all tasks, we will conduct a permutation test, generating a null distribution from 10^5 random permutations of the coupling between individual pupillary and behavioral data sets. To allow comparison across the different tasks, bias effects in individual tasks will be normalized by their range in the null distribution, with 0 and 1 signifying the lowest and highest mean group effect respectively. We will then compare the actual results with the null distribution to test for a significant difference between the high and low pupil response groups in mean normalized bias effect across all tasks. A significant (two-tailed, $p < 0.05$) difference between participants with high and low mean pupillary response in the average bias across all tasks, and no significant difference between either of these groups and those with a medium pupillary response contradicting a monotonic relationship between pupillary response and bias, will constitute weak support in favor of either the effort or the integration account of biased decision making (depending on the direction of the effect). Strong support for either account will require the aforementioned criteria, as well as that no contradictory significant effect is discovered in one of the individual tasks in isolation, while data from at least two of the tasks show a significant effect that aligns with the overall effect (**Table 1**).

Support for:	
Hypothesis 1	$\mu_{\text{high}} < \mu_{\text{low}}$ AND NOT $(\mu_{\text{medium}} > \mu_{\text{low}}$ OR $\mu_{\text{medium}} < \mu_{\text{high}})$
Hypothesis 2	$\mu_{\text{high}} > \mu_{\text{low}}$ AND NOT $(\mu_{\text{medium}} < \mu_{\text{low}}$ OR $\mu_{\text{medium}} > \mu_{\text{high}})$
Level of support:	
Weak	Holds for biases averaged across the six tasks
Strong	Holds for biases averaged across the six tasks, AND holds separately for at least two individual tasks, AND does not support hypotheses 1 on one task and hypothesis 2 on another

Table 1. Criteria for weak and strong support for effort (hypothesis 1) and integration (hypothesis 2) accounts of decision making biases. μ_{high}, μ_{medium}, μ_{low} indicate mean bias effects for the three terciles of participants, divided according to their mean pupillary response (high, medium, and low).

All of the analyses described above, including the quantification of each individual's biases and pupillary responses, and the comparisons at the group level, will proceed precisely as shown in the Supplementary Code that we provide.

We will also use a modeling approach to test for different types of parametric relationships between pupil response and the normalized bias effects across the whole study sample. The purpose of this complementary analysis is to test whether the relationship between pupillary response and biases evident across participants also manifests *within* participants in the changes that occur from question to question and from task to task. The full model will compute the likelihood of a given bias effect for participant s on question q of task t using the following mixed-effects linear regression model:

$$P(\text{bias effect}|s,t,q) = \mathcal{N}\left(\alpha_s + \alpha_{t,q} + \beta_1 P_{s,t,q} + \beta_2 P_{s,t} + \beta_3 P_s; \sigma_s^2 + \sigma_{t,q}^2\right), \tag{3}$$

where $P_{s,t,q}$ is the z-scored pupil response of participant s on question q of task t, $P_{s,t}$ is the average z-scored pupil response of participant s on task t, P_s is the average z-scored pupil response of participant s across all questions of all tasks, all β's are regression coefficients, α_s and $\alpha_{t,q}$ are participant-specific and question-specific intercepts, and σ_s^2 and $\sigma_{t,q}^2$ are participant-specific and question-specific variance terms. This model will be compared to 7 simpler models each omitting one of the 7 terms that comprise the full model. If one of the simpler model wins the model comparison, further simplifications of that model will be tested in the same manner (i.e., by omitting any of the remaining terms). To examine whether the relationship between pupil response and bias differed by task/question, we will compare each model with additional versions of the same model that include regression coefficients for each task or question. Model comparison will be conducted in terms of how well different models predict and fit the data (see **Model predictions** and **Model comparison** below). A log Bayes factor of 10 or more in favor of a model that includes the question and/or task specific regression terms (β_1 and β_2) as compared to a model that does not include these terms will constitute strong evidence for a within-participant relationship between pupil response and bias.

Model predictions. We will compare the different models by calculating how accurately each model predicts participants' biases. Specifically, we will use a 10-fold cross-validation scheme to fit the model to data from a subset of participants ('training set') and generate predicted biases for the remaining participants ('testing set'). Where the model includes participant-specific terms (e.g., a_s), these terms will be instantiated for the testing set with the mean value fitted to the training set. Model accuracy will be computed as the Pearson correlation between actual and predicted mean biases across participants.

Model fitting. To fit the parameters of the different models to observed participant biases, we will use an importance sampling approach[32]. Specifically, we will sample 10^5 random sets of parameter values from predefined prior distributions. We will then compute the likelihood of observing the biases given each parametrization, and use the computed likelihoods as importance weights to derive the posterior distributions. The number of samples may be increased as needed, and will be judged sufficient only if five independent repetitions of the analysis all yield the same conclusions with regards to the parameter values and the model comparison. To define prior distributions, the model-fitting procedure outlined above will be applied to the pilot data using broad priors (normal distribution prior with mean set to 0 and variance set to 100 for the α and β parameters, and inverse gamma distribution with shape and rate set to 0.01 for the σ^2 parameters). The resulting posterior distributions will serve as prior distributions for the main experiment data.

Model comparison. To compare between pairs of models in terms of how well each model fits participants' biases, we will compute the evidence in favor of each model as the mean likelihood of the model given 10^5 random sets of parameter values drawn from the predefined priors. This sampling-based estimate of model evidence accounts for model complexity since it integrates over the entire parameter space.

Quality checks. To ensure that the collected data are able to test the study's hypothesis, we will require three criteria. First, to ensure the quality of the pupil diameter data, we will require that pupillary responses to oddball stimuli be significantly stronger than responses to the other stimuli in the auditory oddball task. Responses to each stimulus will computed as described above (see **Eye tracking subsection of the method**), and then averaged separately for oddball and non-oddball stimuli for each participant. A one-tailed paired t-test ($\alpha = 0.05$) across participants will be used to determine whether responses to oddballs were indeed stronger. If this is not the case, this will indicate that the pupillary recordings are not sufficiently sensitive even to capture this typically robust effect, or else that participants were no paying attention to the oddballs. In either case, new data will need to be collected with a more accurate eye tracking setup, or clearer instruction and more effective incentivization of participants.

Second, since some of our inferences assume a negative correlation between pupillary responses and baseline pupil diameter, we will require that such anti-correlation be evident across participants in the pupil responses to oddball stimuli across the whole experiment. This anti-correlation will be assessed by computing the Pearson correlation across trials between oddball response and pre-stimulus baseline within each participant. We will then conduct a one-tailed t-test across participants to determine whether the average correlation was indeed smaller than 0 ($\alpha = 0.05$). If this is not the case, we will take similar steps as described above for the first quality check.

Third, in the decision-making tasks, a statistically significant bias needs to be evident in at least one of the participant tertiles, when averaged across all six experimental tasks. To average biases across tasks, biases for each task will be scaled such that 1 corresponds to the standard deviation across participants. Biases will then be averaged for each participant, and a one-tailed t-test across participants ($\alpha = 0.05$) will be used to determine whether biases are indeed larger than zero in each of the participant groups. If biases are not evident in any of the groups, this will indicate that our participant group might not have been sufficiently engaged in the experiment, and thus, new data will need to be collected with more effective incentivization of the participants.

Pilot Data

We tested 44 participants on the six decision-making tasks described above (without the oddball task blocks), while measuring their pupil dilation responses. The tercile of participants with highest pupil responses exhibited significant biases on all 6 tasks, whereas the tercile of participants with lowest pupil responses exhibited significant biases only on the Anchoring task (**Figure 3**). Moreover, we found a significant difference between these two groups in the degree to which their decisions were biased across all tasks ($p_{\text{permutation}} < 0.0005$, permutation test; **Figure 4**). Specifically, participants with high pupillary responses (consistent with low neural gain and broader integration) exhibited the strongest and most consistent biases. These results provide preliminary support for the hypothesis that pupillary responses index general susceptibility to decision making biases. In particular, these results suggest that broader integration of information, induced by low neural gain, may play a key role in the formation of biased decision.

We also separately tested 6 participants on the oddball task. The results confirmed that a sequence of 50 tones is sufficient to elicit a robust pupillary response to oddball stimuli ($t_5 > 3.0$, $p < 0.03$, for all 5 blocks), and that a total of 135 tones is sufficient for an anticorrelation between baseline dimeter and pupillary response to emerge ($r < -0.37$ for all 6 participants, $t_5 = 8.6$, $p < 0.001$). In addition, we found no significant habituation of the pupillary response across blocks (mean linear trend $+0.05 \pm 0.12$, $t_5 = 0.4$, $p = 0.72$).

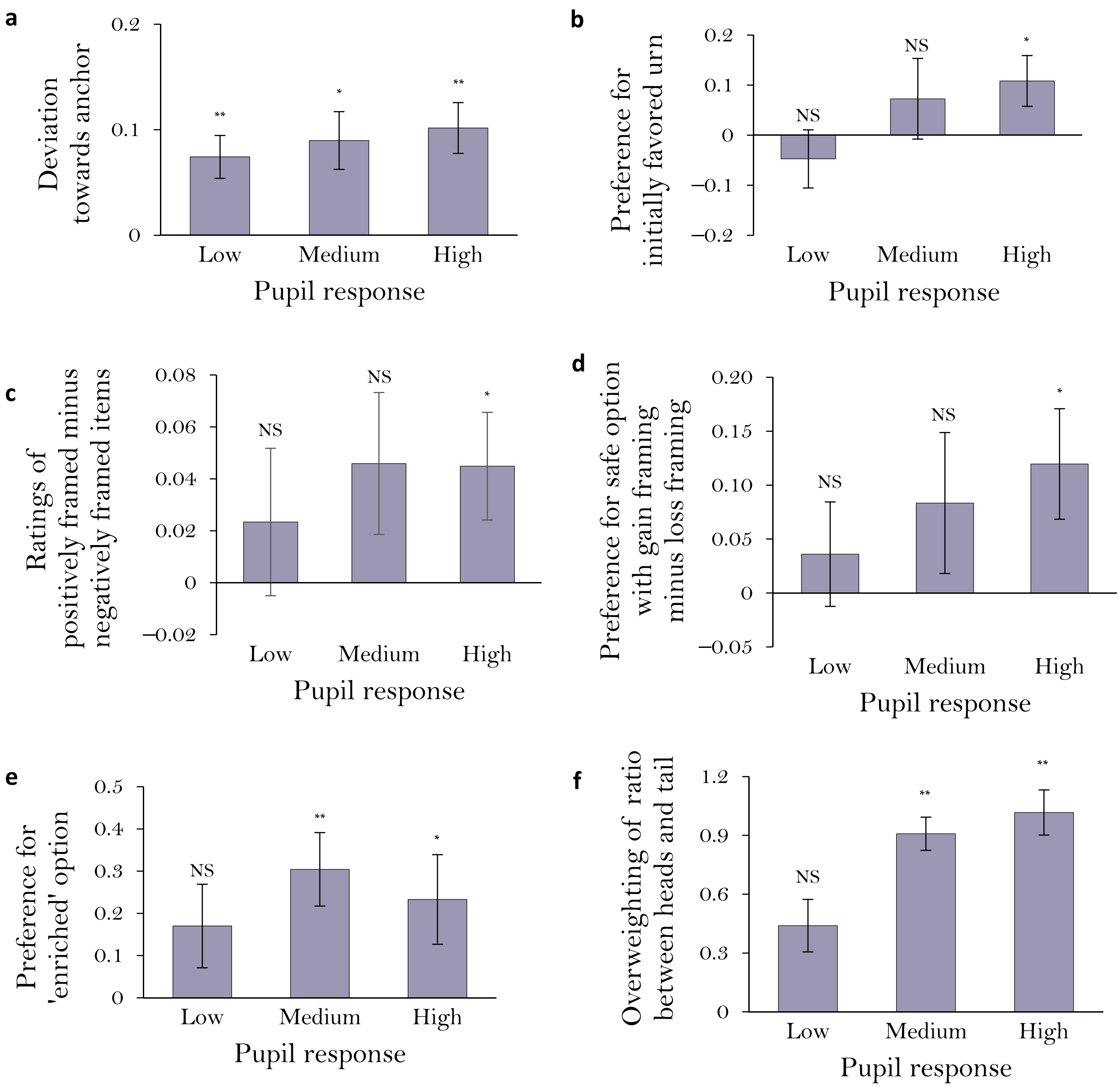

Figure 3. Bias effects in six decision-making tasks as a function of pupil response. For each task, participants were divided into terciles based on mean pupillary dilation in response to task stimuli. Data from between 1 and 9 participants had to be excluded from each task based on the exclusion criteria described in the Methods. NS: $p >$ 0.1, *: $p < 0.01$, **: $p < 0.005$, error bars: across-participant s.e.m. (**a**) **Anchoring.** Deviation of participants' estimates towards the arbitrary anchors they were asked to consider. Estimates were normalized to the range of 0 to 1. n = 40 participants. (**b**) **Persistence of Belief.** Preference of the initially favored urn during the last 60 balls (which were consistent with the other urn). Preferences were indicated on a scale between -1 and 1. An optimal observer would be indifferent on average. n = 35 participants. (**c**) **Attribute Framing.** Difference in evaluation of items framed positively versus negatively. Items were rated on a scale of 0 to 1. Positive values indicate higher evaluations for items framed positively. n = 43 participants. (**d**) **Risky Choice framing.** Increase in risk aversion when outcomes were described in terms of gains as opposed to losses. Preferences were indicated on a scale of -1 to 1. n = 42 participants. (**e**) "**Task Framing".** Preference to both accept and reject the enriched option more than the impoverished option. Preferences were indicated on a scale of -1 to 1. n = 42 participants. (**f**) **Sample-Size Neglect**, measured as the overweighting of the ratio between heads and tails relative to the weight given to the optimal inferences (see Methods). n = 37 participants.

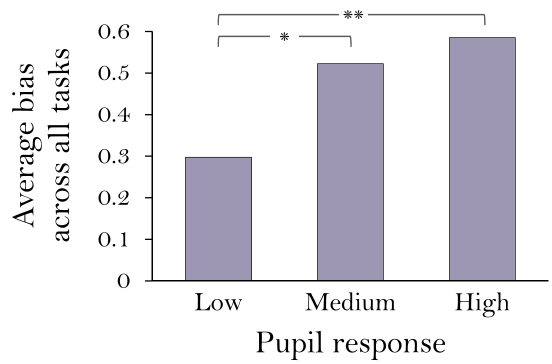

Figure 4. Overall susceptibility to biases. Average normalized bias effect across all tasks. To allow comparison across tasks, individual participant biases for a particular task were scaled and translated such that 0 corresponds to the average bias exhibited by the 1/3 of participants (n=14) with the lowest bias, and 1 corresponds to the average bias exhibited by the most biased 1/3 of participants. Note that terciles were later determined based on pupil response, not bias. n = 44 participants, *: p < 0.01, **: p < 0.0005, permutation test.

Code availability

The custom scripts used for this study are provided as Supplementary Software and are available at https://github.com/eeldar/biases.

Data availability

The data that support the findings of this study are available at https://github.com/eeldar/biases.

Author contribution

Conceptualization, E.E.; Methodology, E.E., V.F., J.D.C and Y.N.; Investigation, E.E. and V.F.; Writing – Original Draft, E.E.; Writing – Review & Editing, E.E., V.F., J.D.C. and Y.N.; Funding Acquisition, Y.N.; Supervision, J.D.C and Y.N.

References

1. Tversky, A. & Kahneman D The framing of decisions and the psychology of choice. *Science* **211**, 453-458 (1981).
2. Kahneman, D. Maps of bounded rationality: Psychology for behavioral economics. *The Am. Econ. Rev.* **93**, 1449-1475 (2003).
3. Fiske, S. T. & Taylor, S. E. *Social cognition: From brains to culture* (Sage, 2013).
4. Usher, M., Tsetsos, K., Erica, C. Y. & Lagnado, D. A. Dynamics of decision-making: from evidence accumulation to preference and belief. *Front. Psychol.* **4**, 758 (2013).

5. Busemeyer, J. R., Jessup, R. K., Johnson, J. G. & Townsend, J. T. Building bridges between neural models and complex decision making behaviour. *Neural Netw.* **19**, 1047-1058 (2006).

6. Krajbich, I. & Rangel, A. A. multi-alternative drift diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl. Acad. Sci. USA* **108**, 13852-13857 (2011).

7. Usher, M. & McClelland, J. L. Loss aversion and inhibition in dynamical models of multialternative choice. *Psychol. Rev.* **111**, 757 (2004).

8. Busemeyer, J. R. & Townsend, J. T. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychol. Rev.* **100**, 432 (1993).

9. Diederich, A. Dynamic stochastic models for decision making under time constraints. *J Math. Psychol.* **41**, 260-274 (1997).

10. Roe, R. M., Busemeyer, J. R. & Townsend, J. T. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychol. Rev.* **108**, 370 (2001).

11. Johnson. J. G. & Busemeyer, J. R. A dynamic, stochastic, computational model of preference reversal phenomena. *Psychol. Rev.* 112, 841 (2005).

12. Kahneman, D. *Attention and effort* (Prentice-Hall, Englewood Cliffs, NJ, 1973).

13. Eldar, E., Cohen, J. D. & Niv, Y. The effects of neural gain on attention and learning. *Nat. Neurosci.* **16**, 1146-1153 (2013).

14. Eldar, E., Niv, Y. & Cohen, J. D. Do you see the forest or the tree? Neural gain and breadth versus focus in perceptual processing. *Psychol. Sci.* **27**, 1632–1643 (2016).

15. Eldar, E., Cohen, J. D., & Niv, Y.. Amplified selectivity in cognitive processing implements the neural gain model of norepinephrine function. *Behav. Brain Sci.* **39**, e206 (2016).

16. Joshi, S., Li, Y., Kalwani, R. M. & Gold, J. I. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* **89**, 221-234 (2016).

17. Servan-Schreiber, D., Printz, H. & Cohen, J. D. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science* **249**, 892-895 (1990).

18. Aston-Jones G. & Cohen J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403-450 (2005).

19. Murphy, P. R., Robertson, I. H., Balsters, J. H. & O'connell, R. G. Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiol.* **48**, 1532-1543 (2011).

20. Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M. & Cohen, J. D. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cogn. Affect Behav. Neurosci. **10**, 252–269 (2010).

21. Reas, C. & Fry, B. *Processing: a programming handbook for visual designers and artists* (Mit Press, 2007).

22. Lambert, A., Wells, I. & Kean, M. Do isoluminant color changes capture attention? *Atten. Percept. Psychophys.* ***65***, 495-507 (2003).

23. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433-436 (1997).

24. Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124-1131 (1974).

25. Jacowitz, K. E. & Kahneman, D. Measures of anchoring in estimation tasks. *Pers. Soc. Psychol. Bull.* 21, 1161-1166 (1995).

26. Peterson, C. R. & DuCharme, W. M. A primacy effect in subjective probability revision. *J. Exp. Psychol.* **73**, 61 (1967).

27. Levin, I. P., Johnson, R. D., Russo, C. P. & Deldin, P. J. Framing effects in judgment tasks with varying amounts of information. *Organ. Behav. Hum. Decis. Process.* 36, 362-377 (1985).

28. Van Schie, E. C. & Van Der Pligt, J. Influencing risk preference in decision making: The effects of framing and salience. *Organ. Behav. Hum. Decis. Process.* **63**, 264-275 (1995).

29. Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263-292.

30. Shafir, E. Choosing versus rejecting: Why some options are both better and worse than others. *Mem. Cognit.* **21**, 546-556 (1993).

31. Griffin, D. & Tversky, A. The weighing of evidence and the determinants of confidence. *Cogn. Psychol.* **24**, 411-435 (1992).

32. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

Acknowledgments

This project was made possible through grants from the Howard Hughes Medical Institute (EE), the Army Research Office (YN), and the Templeton Foundation (VF, JDC, YN).

Competing Interests statement

The authors declare no competing interests.