nature human behaviour

REGISTERED REPORT

https://doi.org/10.1038/s41562-020-01006-3

Check for updates

A pupillary index of susceptibility to decision biases

Eran Eldar[1,2,6] ✉, Valkyrie Felso[3,6], Jonathan D. Cohen[4,5] and Yael Niv[4,5]

The demonstration that human decision-making can systematically violate the laws of rationality has had a wide impact on behavioural sciences. In this study, we use a pupillary index to adjudicate between two existing hypotheses about how irrational biases emerge: 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 is associated with larger pupillary responses. Thus, we tested the relationship between pupil response and choice behaviour on six different foundational decision-making tasks that are classically used to demonstrate irrational biases. Participants demonstrated the expected systematic biases and their pupillary measurements satisfied pre-specified quality checks. Planned analyses returned inconclusive results, but exploratory examination of the data revealed an association between high pupillary responses and biased decisions. The findings provide preliminary support for the hypothesis that biases arise from gradual information integration.

Protocol registration

The stage 1 protocol for this Registered Report was accepted in principle on 19 December 2018. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.c.4368452.v1.

In certain well-described scenarios, human decision-making exhibits systematic deviations from rational behaviour. 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 (for example, the framing effect[1]). The discovery and characterization of such biases has had substantial impact on the fields of psychology and behavioural economics[2]. However, the mechanisms underlying biased decision-making remain widely debated.

The dominant theory 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 that they employ more extensive integration (see Supplementary Discussion and Supplementary Fig. 1 for an example of a computational model illustrating the latter mechanism).

Critically, these two explanatory factors (that is, 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]. In contrast, recent studies have shown 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 behavioural results explained by a hypothesized relationship between high pupillary responses, lower levels of sustained locus coeruleus–norepinephrine function and low neural gain[13,16–26]. In previous theoretical work, we simulated low levels of gain (meaning 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, analysing decision makers' pupil diameter could tell us which mechanism—an automatic, effortless process or extensive integration—is likely to be responsible for generating biased decisions. Furthermore, 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 of 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

[1]Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel. [2]Department of Cognitive Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany. [4]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [5]Psychology Department, Princeton University, Princeton, NJ, USA. [6]These authors contributed equally: Eran Eldar, Valkyrie Felso. ✉e-mail: eran.eldar@mail.huji.ac.il

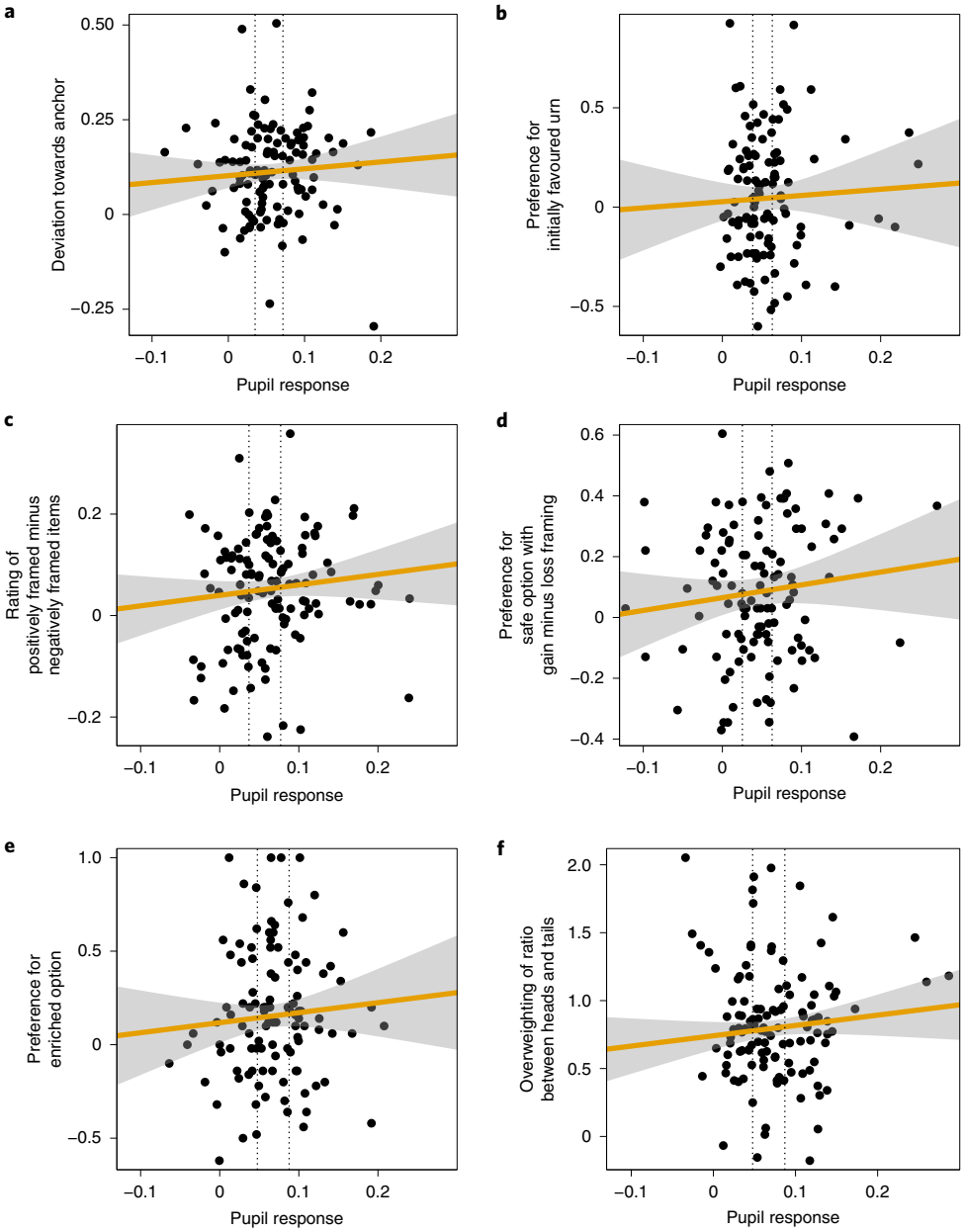

Fig. 1 | Bias effects in six decision-making tasks as a function of pupil response. a, Anchoring task: deviation of participants' estimates towards the arbitrary anchors they were asked to consider. Estimates were normalized to the range 0 to 1 ($n = 112$ participants). **b**, Persistence-of-belief task: preference of the initially favoured urn during the last 60 balls (which were consistent with the other urn). Preferences were indicated on a scale between −1 and 1. An ideal observer would be indifferent on average ($n = 110$ participants). **c**, Attribute-framing task: 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 ($n = 123$ participants). **d**, Risky choice-framing task: 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 ($n = 113$ participants). **e**, Task-framing task: preference to both accept and reject the enriched option more than the impoverished option. Preferences were indicated on a scale of −1 to 1 ($n = 110$ participants). **f**, Sample size neglect task: measured as the overweighting of the ratio between heads and tails relative to the weight given to the optimal inferences (see Methods) ($n = 120$ participants). Each data point represents a participant. Data from 1–9 participants had to be excluded from each task based on the exclusion criteria described in the Methods. The dotted lines divide participants into terciles based on their mean pupillary response to task stimuli. Participants' pupillary responses in different tasks were positively correlated (mean Spearman's $r = 0.48$; range = 0.19 to 0.65). The yellow lines show a robust linear trend, with 95% confidence intervals shown in grey.

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. Of equal importance, the latter result would suggest a potential role for low levels of 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 included 1 min of a classic oddball task between every two test

Table 1 | Biases and their relationship with pupillary and reaction time measures

Bias	Average bias					Task pupil response effect		Oddball pupil response effect		Oddball reaction time effect	
	t	*P*	*M*	Cohen's *d*	95% CI	*P*	$\Delta\mu_{high-low}$	*P*	$\Delta\mu_{high-low}$	*P*	$\Delta\mu_{high-low}$
Anchoring	$t(111)=10.9626$	<0.001*	1.0359	1.0359	0.8045 to 1.2641	0.6152	0.0930	0.4838	0.1255	0.9712	0.0068
Persistence of belief	$t(109)=1.8716$	0.0320*	0.1784	0.1784	−0.0103 to 0.3664	0.6425	0.0852	0.9151	−0.0171	0.8893	−0.0268
Attribute framing	$t(122)=4.9626$	<0.001*	0.4475	0.4475	0.2612 to 0.6320	0.1810	0.2594	0.1009	0.3044	0.1203	−0.3028
Risky choice framing	$t(112)=3.9670$	<0.001*	0.3732	0.3732	0.1817 to 0.5631	0.0992	0.3075	0.4370	0.1358	0.1507	0.2770
Task framing	$t(109)=5.0315$	<0.001*	0.4797	0.4797	0.2813 to 0.6761	0.7174	0.0664	0.8686	0.0307	0.5240	0.1200
Sample size neglect	$t(119)=21.4047$	<0.001*	1.9540	1.9540	1.6467 to 2.2582	0.5352	−0.1208	0.5289	0.1079	0.8668	0.0307

Average biases were tested for significance using two-tailed *t*-tests. Relationships with pupillary or reaction time measures were quantified by dividing participants into terciles based on either pupillary response or reaction time and then comparing the average bias in the low and high terciles. Differences were tested using permutation tests (see Methods for further details). *$P<0.05$.

tasks. The reliable dilation of the pupil in response to oddballs[19,20] served as a positive control. Furthermore, response times on such perceptual discrimination tasks can be expected to reflect neural gain, as indicated by computational modelling 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.

Results

Pre-registered analyses. *Participants demonstrated the expected decision biases and pupillary responses.* Between 110 and 123 participants completed each decision task (see power analysis in Supplementary Fig. 2). First, we examined whether our decision-making tasks were successful in eliciting the expected biases. Participants' behaviour was indeed characterized by systematic biases that were consistent with those observed in previous studies (Fig. 1). Thus, all three participant terciles (divided based on mean pupillary response) showed a significant average bias effect across the six decision tasks (low: $t(41)=9.157$; $P<0.001$; mean (M) = 0.5714; Cohen's $d=1.413$; 95% confidence interval (CI) = 0.9791 to 1.8379; medium: $t(41)=11.5237$; $P<0.001$; $M=0.8581$; Cohen's $d=1.7781$; 95% CI = 1.2853 to 2.2622; high: $t(41)=10.812$; $P<0.001$; $M=0.8393$; Cohen's $d=1.6683$; 95% CI = 1.1938 to 2.1340; all one-tailed *t*-tests). Furthermore, a bias effect robustly manifested in all of the individual tasks (Table 1).

To assess the validity of our pupillometry measurements, we next tested for the anticipated oddball effect on pupil dilation. Examination of the timecourses of pupillary response confirmed that our pre-specified time windows were appropriate for capturing responses on decision and oddball trials (Supplementary Fig. 3). As expected, we found that responses to oddball tones ($M=0.0782$) were significantly higher than responses to non-oddball tones ($M=0.0362$) ($t(122)=18.809$; $P<0.001$; $M=0.042$; Cohen's $d=1.6959$; 95% CI = 1.4180 to 1.9709; one-tailed *t*-test). In addition, we found the expected anti-correlation between pupillary responses and baseline pupil diameter ($t(122)=-21.0411$; $P<0.001$; $M=-0.4734$; Cohen's $d=-1.8972$; 95% CI = −2.1920 to −1.5995; one-tailed *t*-test), consistent with our previous work[13,14].

Comparison of participant pupil diameter terciles did not support either hypothesis. Next, we compared participants with low and high pupillary responses in terms of how biased they were across the six decision tasks. Although a monotonic increase manifested across terciles, the difference between the high and low terciles was not significant (overall permutation test: $P=0.143$; $\Delta\mu_{high-low}=0.1151$; Table 1 and Fig. 2a).

As alternative measures of participants' physiological state, we also examined pupillary responses and reaction times in the oddball tasks that preceded and followed each decision-making task. Participants' pupillary responses to oddball stimuli correlated with their pupillary responses in the decision-making tasks ($r=0.43$; $P<0.001$). This correlation provides important confirmation that pupillary responses in the decision tasks indeed reflected individual differences that manifested similarly across two very different types of task. However, dividing participants into terciles based on pupillary responses to oddballs yielded similar non-significant results with regards to decision biases (overall permutation test: $P=0.1247$; $\Delta\mu_{high-low}=0.1145$; Table 1 and Fig. 2b).

Reaction times to oddball stimuli did not consistently correlate with pupillary responses in the oddball task ($r=-0.18$; $P=0.0522$), nor did they correlate with pupillary responses in the decision-making tasks ($r=-0.01$; $P=0.9457$), and dividing participants into terciles based on these reaction times also resulted in non-significant results (overall permutation test: $P=0.8306$; $\Delta\mu_{high-low}=0.0175$; Table 1 and Fig. 2c).

Planned trial-level analysis was found to be infeasible. Following data collection, we discovered that the planned modelling analysis could not be applied to the full dataset, since the large number of free parameters (for example, an intercept for each participant) created too many possible combinations of parameter values. Consequently, the results obtained using importance sampling did not replicate on repeated execution of the analysis, even when the number of samples was increased to the maximum number supported by our computing resources (10^7). Therefore, we implemented a trial-by-trial model using two alternative complementary approaches: hierarchical Bayesian inference (to determine the posterior distribution of model parameters) and frequentist linear mixed-effects model fitting and comparison. Both of these alternative approaches mitigate the difficulties that arise due to large numbers of free parameters by sampling parameters from adaptive group-level prior distributions that restrict individual-level parameter flexibility.

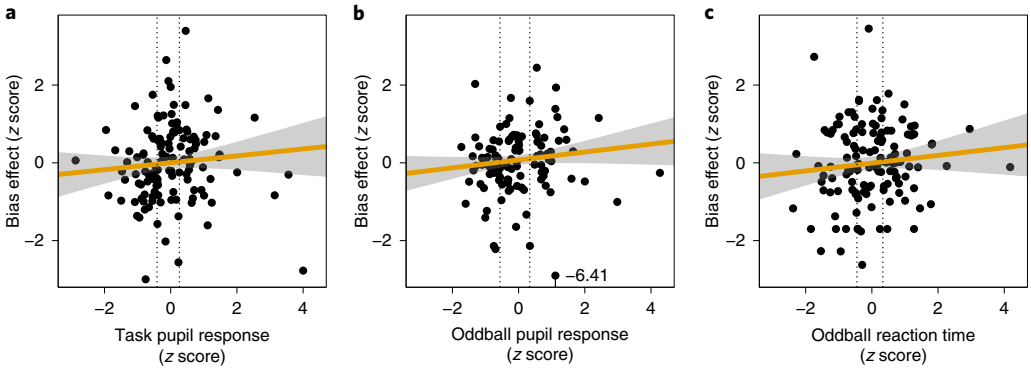

Fig. 2 | Overall susceptibility to biases (average normalized bias effect across all tasks). a, Biases as a function of pupil response during the tasks ($n = 126$ participants). **b**, Biases as a function of pupil response in a standardized oddball task ($n = 112$ participants). **c**, Biases as a function of reaction time in a standardized oddball task ($n = 132$ participants). The dotted lines divide participants into terciles based on mean pupillary response (**a** and **b**) or reaction time (**c**). The yellow lines show robust linear trends, with 95% confidence intervals shown in grey.

Table 2 | Bayesian inference concerning the relationship between pupillary responses and decision biases

Bias	Participant-level pupillary effect			Task-level pupillary effect			Question-level pupillary effect			Total pupillary effect		
	β_3	95% CI	P_d	β_2	95% CI	P_d	β_1	95% CI	P_d	$\Sigma\beta$	95% CI	P_d
Anchoring	0.0246	−0.0251 to 0.0710	0.8307	0.0639*	0.0082 to 0.1198	0.9882	−0.0138	−0.0671 to 0.0371	0.2896	**0.0681**	**−0.0138 to 0.1567**	**0.9519**
Persistence of belief	0.0073	−0.0265 to 0.0451	0.7101	0.0316*	0.0003 to 0.0597	0.9765	−0.0023	−0.0289 to 0.0223	0.4294	**0.0386**	**−0.0222 to 0.0965**	**0.8983**
Attribute framing	0.0705*	0.0124 to 0.1277	0.9924	0.0126	−0.0493 to 0.0764	0.6514	0.0289	−0.0213 to 0.0832	0.8731	**0.1123***	**0.0192 to 0.2120**	**0.9886**
Risky choice framing	0.1128*	0.0093 to 0.2358	0.9799	0.0434	−0.0534 to 0.1372	0.7843	0.02617	−0.1082 to 0.1565	0.6638	**0.1855***	**0.0140 to 0.3614**	**0.9769**
Task framing	0.0370	−0.0319 to 0.1226	0.8618	0.0271	−0.0438 to 0.1090	0.8082	0.0726*	0.0011 to 0.1601	0.9759	**0.1561***	**0.0295 to 0.2842**	**0.9897**
Sample size neglect	−0.0298	−0.0826 to 0.0247	0.1486	0.0032	−0.0481 to 0.0546	0.5555	0.0213	−0.0199 to 0.0648	0.8345	**−0.0012**	**−0.0797 to 0.0751**	**0.4700**
Main effect across tasks	**0.0365***	**0.0065 to 0.0730**	**0.9890**	**0.0323***	**0.0026 to 0.0578**	**0.9840**	**0.0246**	**−0.0075 to 0.0547**	**0.9285**	**0.0925***	**0.0418 to 0.1420**	**0.9999**

The participant-level effect reflects variance among participants in mean pupillary response. The task-level effect reflects within-participant variance in mean pupillary response for different tasks. The question-level effect reflects within-participant, within-task variance in pupillary response for different questions. The total pupillary effect ($\Sigma\beta$) reflects the cumulative impact of all sources of variance on pupil response. See 'Hierarchical Bayesian modelling' in the Methods for details. In this table, CI denotes credible interval. Coeffecients were estimated by the mode of their posterior distribution. *$P_d > 0.975$.

Exploratory analyses. *Trial-level analysis showed that biases were associated with stronger pupillary responses.* Examining participants' biases on a trial-by-trial basis using hierarchical Bayesian modelling revealed a significant main effect of pupillary response on the extent of the bias (Table 2). This effect was significant for participants' mean pupillary response (β_3 in the Table 2), as well as for the differences in pupillary responses across tasks for each participant (β_2). A preponderance of evidence (probability of direction (P_d) = 0.9285) also supported a trial-level effect within tasks (β_1). Examining each task individually showed significant pupillary effects on each of the three framing biases (Fig. 3a). Similar pupillary effects on anchoring and persistence-of-belief (Supplementary Fig. 4) biases also received some support ($P_d > 0.89$). Furthermore, when comparing between pairs of tasks, we found significant differences in pupillary effects in only two out of 15 possible comparisons (Supplementary Table 1). These results suggest a reasonably consistent general relationship between pupillary response and bias.

Model comparison between different linear mixed-effects models also revealed a significant relationship between behavioural biases and pupillary responses, specifically with regards to participant mean pupillary response (β_3). Thus, the model that best explained trial-by-trial biases included a fixed effect for this predictor as well as a random slope that varied across questions (Bayesian information criterion (BIC) = 11,673.88). This model outperformed a similar model without pupillary predictors (BIC = 11,691.10) by an estimated log[Bayes factor] of 8.61. The pupillary coefficient indicated a positive relationship between pupillary response and behavioural bias ($\beta_3 = 0.0649$; s.e. = 0.0299; $t(62.4) = 2.171$; $P = 0.0337$). Finally, a similar analysis also showed a positive relationship between bias in the decision-making tasks and pupillary response to the interleaved oddball stimuli ($\beta_3 = 0.0279$; s.e. = 0.0299; $t(67.2) = 2.084$; $P = 0.0410$).

Given the discrepancy between the tercile and modelling analyses, we inquired whether this difference reflected the former analysis' reduced sensitivity due to the binning of participants and/or the latter analysis' increased sensitivity due to standardization of biases by question. To test this, we first averaged the bias effects that were quantified and standardized per question for each participant tercile. We found a significant difference between participants with low and high pupillary response (overall permutation test: $P = 0.0040$;

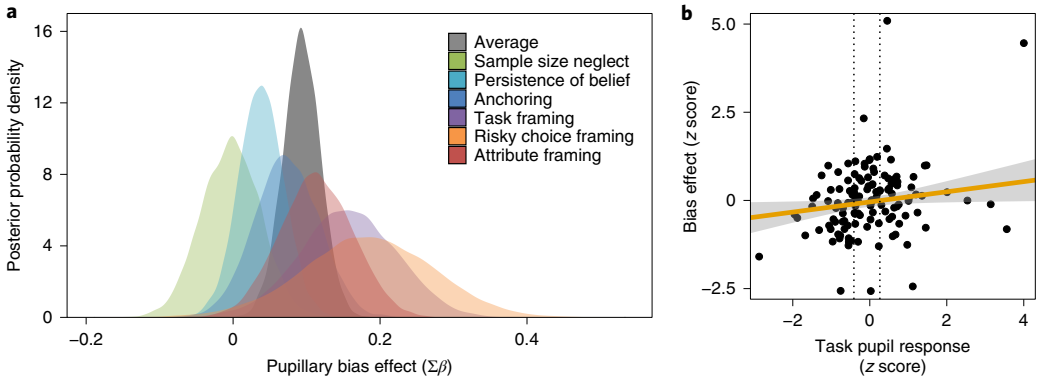

Fig. 3 | Trial-level analysis of biases. a, Posterior probability of pupillary effects ($\beta_1+\beta_2+\beta_3$) on biases exhibited in each decision-making task, as inferred using hierarchical Bayesian modelling. Each density was constructed from 10,000 Monte Carlo Markov chain samples. The average effect is significant, although the effects for each of the tasks in isolation are significant only for the three framing tasks. **b**, Bias effects, standardized for each question and then averaged over questions, as a function of pupillary response in the decision-making tasks. Biases were quantified separately for each question and *z* scored across participants ($n=126$ participants). The dotted lines divide participants into terciles based on pupillary response. The yellow line shows a robust linear trend, with the 95% confidence interval shown in grey.

Table 3 | Criteria for weak and strong support for effort (hypothesis 1) and integration (hypothesis 2) accounts of decision-making biases

Conclusion	Criteria
Support for hypothesis 1	$\mu_{high} < \mu_{low}$ and not ($\mu_{medium} > \mu_{low}$ or $\mu_{medium} < \mu_{high}$)
Support for hypothesis 2	$\mu_{high} > \mu_{low}$ and not ($\mu_{medium} < \mu_{low}$ or $\mu_{medium} > \mu_{high}$)
Weak level of support	Holds for biases averaged across the six tasks
Strong level of support	Holds for biases averaged across the six tasks, holds separately for at least two individual tasks and does not support hypothesis 1 on one task and hypothesis 2 on another

μ_{high}, μ_{medium} and μ_{low} indicate mean bias effects for the three terciles of participants, divided according to their mean pupillary response (high, medium and low, respectively).

$\Delta\mu_{high-low}=0.2452$; Fig. 3b). At the same time, performing the trial-level mixed modelling with biases standardized by task (as in the original pre-registered analysis) also yielded a significant positive relationship between bias and pupillary response ($\beta_3=0.0556$; s.e. $=0.0218$; $t(61.9)=2.551$; $P=0.0132$).

Discussion

We replicated a set of systematic decision-making biases and examined their relationship with pupillary responses in order to differentiate between two competing hypotheses for the source of the biases: if biases are due to fast, effortless (and therefore sloppy) processing of information, they should correlate with smaller pupillary responses, as pupil diameter is an established index of processing effort[12]. Alternatively, if biases result from extensive integration of evidence that brings information that should not affect a decision to bear on that decision[4–11], they should correlate with larger pupillary responses, indicating lower neural gain and more expansive information processing[13–20].

Our main planned analyses could not adjudicate between the two hypotheses conclusively (Table 3). However, trial-level quantification of bias effects showed that high pupillary responses were associated with stronger biases. Although this analysis was also planned, the details of its implementation had to be modified from the original plan, due to infeasibility of the pre-specified approach. We therefore consider it prudent to regard these results as exploratory. Thus, we conclude that the results provide preliminary support for the association of high pupillary responses with susceptibility to decision biases and more decisively contradict the opposite possibility (namely, that biases are associated with lower pupillary responses).

The idea that decision biases reflect the operation of an effortless automatic system, and can be over-ridden with effortful deliberation, pervades the scientific and popular literature[21]. However, this notion is contradicted by a body of literature showing that monetary incentives do not eliminate biases[22]. Indeed, in the present study, lower pupillary responses, which suggest lower effort, were not significantly associated with more biased decisions.

The preliminary evidence that we do find in favour of the opposite relationship joins together two disparate lines of research. The first suggests that higher pupillary responses are associated with low baseline pupil diameter, and thus with low tonic norepinephrine function[18], which slows down and broadens the integration of information that gives rise to decisions[13–15]. The second suggests that gradual integration of information is essential for biases to emerge[4,5]. Together, these two lines of research predict that low norepinephric tone, which is indicated by high pupillary responses, will be associated with higher susceptibility to decision biases.

While our results fall short of providing conclusive support for this hypothesis, they present several pieces of evidence in its favour. The first, just discussed, is the positive association between pupillary responses and biases in the trial-level analyses. Second, we found few differences between decision-making tasks in the relationship between pupillary responses and biases, suggesting that this relationship is not specific to only one type of task. Third, the index of pupillary response generalized to some degree across decision-making and oddball tasks, as would be expected from an index of individual differences in general neuromodulatory state.

Our preliminary results might also seem to contradict a body of work originating in the field of perceptual decision-making that observed an association between decision biases and low pupillary responses[23–26]. However, this apparent discrepancy can be resolved by considering that (different from our tasks) bias, as it is defined in perceptual decision-making, is not necessarily irrational. For instance, a participant is considered to exhibit a perceptual decision-making bias if they are more likely to say that a visual

stimulus was not presented when it was than they are to say that a stimulus was presented when it wasn't, even if they have valid reason to prefer one type of mistake over the other (for example, the participant might not want to seem like someone who is seeing things). The possible rationality of this type of bias is underscored by a recent study showing that the bias is sensibly adjusted to the overall probability of a stimulus being presented[26] (that is, participants become more likely to err by saying a stimulus was presented if the proportion of trials where stimuli are presented is increased). In contrast, the present study examined biases that have been regarded as classic examples of irrational decision-making.

From a mechanistic perspective, we propose that irrational biases in particular often arise from a gradual process of integration that allows weak irrelevant influences to impact the decision. This proposal is based on combining the present results with previous findings that high pupillary responses are associated with broader, more gradual integration of evidence[13,14]. Moreover, computational modelling and consistent functional magnetic resonance imaging findings[13] further suggest that this link between high pupil response and broad integration may be mediated by low tonic norepinephrine function. We note though that high pupil responses could also simply reflect a high phasic norepinephrine response[16]. It is likely that these two sources of variance (the baseline diameter and the phasic norepinephrine response) have different relative contributions to the pupillary response in different experimental models and thus inferences about underlying neuromodulators need to be made with caution. Thus, further work is needed to determine whether tonic low norepinephrine gives rise to susceptibility to decision-making biases.

In summary, our results provide preliminary evidence in support of an association between high pupillary responses and decision-making biases. This finding disagrees with the explanation of susceptibility to biases as reflecting decreased effort or automatic processes, and is instead consistent with the hypothesis that biases arise from a gradual process of information integration that occurs under low levels of neural gain.

Methods

Participants. A target sample size of 120 participants was determined via a bootstrapping-based power analysis of pilot data (see 'Power analysis' below). To achieve this sample size, we recruited 159 participants from the greater Princeton area (mean age = 19.9 years; age range = 18–33 years; 105 female; 54 male). Inclusion criteria were an age of 18–35 years and compatibility with pupillometry, as evidenced by successful calibration of the eye tracker. Participants gave written informed consent before taking part in the study, which was approved by the Princeton University Institutional Review Board. Participants received either course credit (91/159) or compensation of $12 per hour (68/159) for participation. Data from 19 participants were not analysed due to unsuccessful calibration of the eye tracker. Examination of the eye tracking data according to our a priori exclusion criteria (see 'Eye tracking' below) revealed that eight participants had no valid data for any of the tasks and five additional participants had no valid data for the oddball task. In addition, five participants were removed due to a programming error that prevented the tasks from running correctly, and one participant was removed due to having previously performed the experiment. Therefore, in total, we tested 121 participants with valid task and oddball data.

Power analysis. To determine the target 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 1,000 datasets, each of which was constructed by sampling participants with replacement from the pilot data. The power analysis showed that a sample size of 120 participants provided a 95% probability of finding strong support for the study's hypothesis, given the effect size found in the pilot data (Supplementary Fig. 2). 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.

Stimuli. Stimuli were generated using the Processing programming environment[27]. To minimize luminance-related changes in pupil diameter, we first identified colours that were isoluminant with the background by having each participant perform the flicker fusion procedure[28]. The colours of the experimental stimuli were then automatically adjusted accordingly, to achieve subjective isoluminance in the conditions of the testing room, for each participant. Stimuli were presented on a computer screen using MATLAB software (MathWorks) and the Psychophysics Toolbox[29].

Experimental design. Each participant performed six experimental tasks, each aimed at inducing a different bias. To facilitate comparisons between participants, all participants performed tasks in the order in which the tasks are described below. Each experimental session lasted approximately 1 h (this varied due to different calibration durations and the self-paced nature of the tasks). Unless otherwise noted, questions appeared on the screen until the participant entered their answer using a keyboard (that is, there were no time restrictions for providing an answer). To allow sufficient time for pupillary responses to be resolved, questions were separated by random inter-trial intervals (7–9 s long; uniformly distributed), during which only a fixation cross appeared on the screen.

Anchoring task. For the anchoring task[30], participants answered two questions about each of seven quantities (for example, the height of the Eiffel tower). They were first asked to indicate whether the quantity is greater ('1' keyboard key) or smaller ('2' keyboard key) than an anchor value. Once the participant responded, the first question disappeared from the screen and the participant was immediately asked to estimate the quantity by typing it using the keyboard and then pressing ENTER. Each quantity was coupled with a low anchor for half of the participants and with a high anchor for the other half. Each participant was presented with a low anchor for half (three or four) of the quantities and with a high anchor for the other half. Quantities and calibrated anchor values were taken from a previous study[31], including: length of the Mississippi River; population of Chicago; number of babies born per day in the United States; 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 were 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 was therefore quantified by the deviation of an estimate in the direction of the anchor relative to the mean estimate provided by the whole study sample. Three estimates whose distance from all other participants' mean estimate was more than ten times the range of the other participants' estimates were excluded as outliers.

Persistence-of-belief task. For the persistence-of-belief task[32], participants were presented with two urns, each filled with ten coloured balls (Supplementary Fig. 4a). One urn contained three red balls, two green balls, two blue balls, two brown balls and one purple ball, while the other urn contained two red balls, three green balls, one blue ball, two brown balls and two purple balls. Participants were then shown a sequence of 90 balls, which they were told were sampled with replacement from one of the urns. Each sampled ball fell from the top of the screen, horizontally centred, until it settled near the bottom of the screen and then disappeared. Balls followed one another in sequence without a break (3.3 s per ball) while the two urns were presented on the left and right sides of the screen. Every five samples (balls), participants were asked to indicate using an appropriately labelled horizontal sliding bar which urn they thought the sequence was sampled from. Participants were instructed to indicate their degree of certainty by means of the precise position of the bar, where a central position corresponded to total uncertainty. Each question was followed by an inter-trial interval. The sequence of balls was set up so that the first 30 balls favoured one of the urns as their source with a probability of 0.95 and the next 60 balls favoured the other urn to a similar degree (per 30 balls). Therefore, it was optimal to favour one urn after 30 balls, be indifferent after 60 balls and favour the second urn after 90 balls (Supplementary Fig. 4b). 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 was therefore quantified as the degree to which each participant's average response during the last 60 balls favoured the initially favoured urn. Similarly, for trial-level analyses, the negative of a persistence-of-belief effect was quantified as the degree to which the participant updated their preference towards the second-favoured urn, minus the participant's average update in the opposite direction during the first 30 trials. (Note that the stage 1 protocol did not specify precisely how we would quantify this bias on a trial-by-trial basis.) The initially favoured urn was counterbalanced across participants. Data from ten participants who did not favour the correct urn during the first 30 balls were excluded from the analysis.

Attribute-framing task. For the attribute-framing task[33], participants were asked to rate ground beef products, gambles and students' performance, whose attributes were framed either positively or negatively. In the ground beef task, participants were asked to imagine they were having a friend over for dinner and they were

about to make their favourite lasagne dish with ground beef. They were then asked to rate how satisfied they would be purchasing each of four ground beef products, described in terms of price per pound ($2.7 or $3.3) and either percentage lean (80 or 90%; positive frame) or percentage fat (20 or 10%; negative frame). In the gambles task, participants were asked to imagine that they had $10 and could either keep the $10 or pay the $10 to take a gamble. They were then asked to rate how likely they were to take each of three gambles, described in terms of amount to be won ($50, $100 or $200) and either probability of wining (20, 10 or 5%; positive frame) or probability of losing (80, 90 or 95%; negative frame). In the student performance task, participants were asked to evaluate each of two students on the basis of midterm exam and final exam performance, described in terms of either percentage correct (50 or 70%; positive frame) or percentage incorrect (50 or 30%; negative frame). The attributes of an item remained on the screen until the participant finished rating the item by adjusting an appropriately labelled vertical sliding bar and then pressing ENTER. Each item was framed positively for half of the participants and negatively for the other half. For a given participant, all items of a particular type were similarly framed (that is, either positively or negatively) so as to minimize awareness of the framing manipulation, but the framing was varied within participants across item types. As in the anchoring task, the framing effect was quantified for each item by the deviation of a participant's rating from the overall mean rating, in the direction of the frame (that is, upwards for positive frames and downwards for negative frames).

Risky choice-framing task. For the risky choice-framing task[34], participants faced two different scenarios (a medical scenario and a fire scenario) and were asked to indicate using a sliding bar which of two available actions they would choose in each scenario. One action had a certain outcome and the other had an uncertain outcome, both of which were framed in terms of either gains or losses (counterbalanced across participants). The scenarios were described in full, as was done previously[34]. In the medical scenario, which concerned the treatment of a deadly disease on an island inhabited with 600 inhabitants, participants were asked to choose between the 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 concerned the treatment of fires threatening 9,000 acres of forest, participants were asked to choose between gain-framed outcomes ('3,000 acres of forest will be saved' and 'a 60% chance that 5,000 acres will be saved and a 40% chance that no forest under threat will be saved') or between loss-framed outcomes ('6,000 acres of forest will be lost' and 'a 60% chance that 4,000 acres will be lost and a 40% chance that 9,000 acres will be lost'). For each question, the attributes of the first option (as described above) appeared on the left side of the screen and the attributes of the second option appeared on the right side of the screen. These details remained on the screen until the participant indicated their preference by adjusting an appropriately labelled horizontal sliding bar and then pressing ENTER. As for the anchoring and attribute-framing tasks, the framing effect was quantified as the deviation of a participant's preference from the overall mean rating, in the direction of the frame (that is, 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[35]).

Task-framing task. For the task-framing task[36], participants faced five different problems concerning various subjects, such as child custody, vacation choice, ice-cream choice and gambling. Each problem involved one option with more positive and negative attributes (the enriched option) and one option with fewer positive and negative attributes (the impoverished option). In each problem, half of the participants were asked to choose one of the two options and the other half were asked to reject one of the two options. For example, in one problem, participants were asked to imagine that they served on the jury of an only-child sole-custody case following a relatively messy divorce and they had to make a decision based entirely on the following few observations: average income, average health, average working hours, reasonable rapport with the child and relatively stable social life (parent A; no particularly positive or negative attributes); or above-average income, very close relationship with the child, extremely active social life, lots of work-related travel and minor health problems (parent B; three positive attributes and two negative attributes). Half of the participants were asked to which parent they would award sole custody of the child, while the other half were asked which parent they would deny sole custody of the child. A full description of the other problems can be found elsewhere[36] (problems 1, 2, 4, 5 and 6). Participants were 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 versus reject) was varied within participants across questions. The task-framing bias manifests in people's tendency to choose (either award or reject) the enriched option as opposed to the option they have less conclusive 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 (that is, award frame) or reject one option (that is, reject frame). Thus, the framing effect was quantified by the degree to which each participant chose the enriched option (that is, parent B) more frequently than the impoverished option (that is, parent A). In accordance with the participant-level analysis, trial-level framing effects were quantified by the degree to which the participant preferred the enriched option compared with the group average under the same frame. (Note that the stage 1 protocol did not specify precisely how we would quantify this bias on a trial-by-trial basis).

Sample size neglect task. For the sample size neglect task[37], participants were asked to imagine that they were tossing a biased coin and recording how often the coin landed heads and how often the coin landed tails. They knew that the coin was bent and tended to land on one side three out of five times, but they did not know whether this bias was in favour of heads or in favour of tails. Participants were then presented with ten different sets of results (number of heads and number of tails), in which the heads always outnumbered the tails, and they were asked to indicate using a vertical sliding bar how certain they were, given each set, that the coin was biased in favour of heads. The top end of the bar was labelled with "completely certain that coin favours heads" and the bottom end was labelled with "completely uncertain that coin favours heads". Each set of results remained on the screen until the participant finished adjusting the bar and pressed ENTER. The sets of results were similar to those used previously[37].

As shown by Griffin and Tversky[37], the probability that the coin is biased in favour of heads, according to Bayes' rule, is:

$$P(H|D) = e^{(h-t)\log_2^3}, \quad (1)$$

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

$$P(H|D) = e^{n\frac{(h-t)}{n}\log_2^3} = e^{n\frac{(h-t)}{(h+t)}\log_2^3}, \quad (2)$$

which depends on the sample size (that is, the number of outcomes, n) and the observed ratio of heads and tails $\left(\frac{h-t}{h+t}\right)$. Previous work has shown that people tend to overweigh the ratio component at the expense of the sample size component (sample size neglect[37]). Thus, to measure this bias for an individual participant, we regressed the participant's estimates against the true probabilities (equation (1)) as well as against the ratio component $\left(e^{\frac{h-t}{h+t}\log_2^3}\right)$ and compared the two resulting regression coefficients (β_{true} and β_{ratio}). All inputs to the regression analyses were 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 was computed for each participant as $1 - \beta_{\text{true}} + \beta_{\text{ratio}}$. In accordance with the participant-level analysis, trial-level sample size neglect was quantified as the residual obtained by regressing the participant's ratings against the true probabilities minus the residual obtained by regressing ratings against the ratio component. (Note that the stage 1 protocol did not specify precisely how we would quantify this bias on a trial-by-trial basis.)

Data from three participants for whom β_{true} and β_{ratio} were lower than 0, or who reported higher certainty given three heads and two tails than given seven heads and two tails were excluded from the analysis. The former criterion indicates that the participant did not give reasonable answers, whereas the latter criterion suggests specifically that the participant mistakenly looked for a ratio that best matched three to two.

Oddball task. To assess reaction times and pupillary responses in a uniform manner throughout the experiment, and as a positive control to our other findings, we used 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 were presented with a sequence of 60-ms sinusoidal tones of two possible frequencies: 1,000 Hz (which were designated as the target) and 500 Hz (which were designated as non-targets). Participants were told to respond with a keypress only when the target tone was sounded. Inter-tone intervals were drawn uniformly between 2.1 and 2.9 s. To allow the pupil diameter to return to baseline, the stimuli were ordered such that target tones were always spaced between at least three non-target tones on each side. Target tones made up 20% of the tones. The results of pupil diameter response to the oddball items were analysed to verify reliable pupillometry measurements. As in previous studies[19], we excluded from the analysis trials in which a participant responded to a non-target tone (false positive; $M = 2.2$ out of 140 non-target tones per participant), did not respond to a target tone (miss; $M = 1.3$ misses out of 35 targets per participant) or responded within 100 ms of target presentation (quick response; $M = 0.6$ out of 35 per participant).

Participants performed a total of seven oddball task blocks, such that oddball blocks alternated with the six decision-making tasks. Each block consisted of 25 tones (five of them oddballs). Oddball reaction time and pupillary response were computed for each decision-making task based on the oddball blocks that immediately preceded and followed the task (that is, based on a total of 50 tones/ten oddballs). These measures were 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 120 Hz eye tracker (SensoMotoric Instruments) was used to measure participants' left and right pupil diameters at a rate of 120 samples per second while they were performing the behavioural tasks with their head fixed on a chinrest. At the beginning of the experiment, a baseline measurement of pupil diameter at rest was taken for a period of 45 s. Pupil diameter data were analysed in MATLAB, as in previous work[13,14]. First, the data were processed to detect and remove blinks and other artefacts. For this purpose, artefactual diameter samples were identified as those lower than 66% or higher than 150% of the median non-zero sample, as well as those samples that differed from adjacent samples by more than 10%. Samples recorded between 33 ms before an artefact and 100 ms after it were also designated as artefactual. Following data collection, we noted that a small proportion of samples were dropped (that is, the gap between recorded samples was larger than expected), so we treated these dropped samples as artefacts as well. All artefactual samples were replaced by linear interpolation. For each task and each question, the baseline pupil diameter was computed as the average diameter over a period of 1 s before presentation of the question. Based on an examination of the pilot data, we determined that in the six decision-making tasks, the pupil dilation response would be computed as the peak diameter recorded during the period between 1 and 6 s following presentation of the question, minus the preceding baseline diameter. For the oddball task, pupil responses were shorter; thus, the peak diameter was assessed between 0.4 and 2 s following stimulus onset. All pupil dilation responses were normalized by the pre-experiment baseline pupil diameter. Questions and oddball trials for which more than half of the pupil measurements were affected by artefacts were considered invalid and excluded from the analysis. Participants with fewer than two valid (that is, mostly artefact-free) questions in a given task were excluded from the analysis of that task ($M = 6.5$ participants per task).

Statistical analysis. For each task, we divided participants into terciles of low, medium and high mean pupil dilation. This allowed 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 conducted a permutation test, generating a null distribution from 10^5 random permutations of the coupling between individual pupillary and behavioural datasets. To allow comparison across the different tasks, bias effects in individual tasks were normalized by their range in the null distribution, with 0 and 1 signifying the lowest and highest mean group effect, respectively. We then compared 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. Before data collection, we pre-specified that 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, would constitute weak support in favour of either the effort or the integration account of biased decision-making (depending on the direction of the effect). Strong support for either account would require the aforementioned criteria, as well as that no contradictory significant effect was discovered in one of the individual tasks in isolation, while data from at least two of the tasks showed a significant effect that aligned with the overall effect (Table 3).

All of the analyses described above, including the quantification of each individual's biases and pupillary responses, as well as the comparisons at the group level, proceeded precisely as shown in the analysis code available at https://osf.io/sygz3/.

Trial-level modelling. We also used a modelling approach to test for different types of parametric relationship between pupil response and the normalized bias effects across the whole study sample. The primary purpose of this complementary analysis was to test whether the relationship between pupillary response and biases that was evident across participants also manifested within participants in the changes that occurred from trial to trial and from task to task. Since effect size was likely to vary by question, and since questions were administered to all participants in precisely the same order, trial-level effects were each normalized to a common scale by translation and scaling such that 0 corresponded to the average effect and 1 corresponded to the standard deviation of the effects. (Note that the stage 1 protocol did not specify precisely how we would normalize trial-level effects as they could be averaged across questions and tasks.)

The full model computes the likelihood of a given bias effect for participant s on question q of task t using the following mixed-effects linear regression model:

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

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 βs 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 was to be compared with seven simpler models, each omitting one of the seven terms that comprised the full model. If one of the simpler models had won the model comparison, further simplifications of that model would have been tested in the same manner (that is, by omitting any of the remaining terms). To examine whether the relationship between pupil response and bias differed by task or question, we planned to compare each model with additional versions of the same model that included regression coefficients for each task or question. Model comparison was to be conducted in terms of how well different models predicted and fit the data. A log[Bayes factor] of ten or more in favour of a model that includes the question and/or task-specific regression terms (β_1 and β_2) compared with a model that does not include these terms would constitute strong evidence for a within-participant relationship between pupil response and bias.

The planned modelling approach was found to be infeasible, and was replaced by Bayesian and mixed modelling approaches (see 'Exploratory analyses').

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

Model comparison procedures were adapted to the exploratory modelling approaches (see 'Exploratory analyses').

Model fitting. To fit the parameters of the different models to observed participant biases, we planned to use an importance sampling approach[38]. Specifically, we planned to sample 10^5 random sets of parameter values from predefined prior distributions and then compute the likelihood of observing the biases given each parametrization and used the computed likelihoods as importance weights to derive the posterior distributions. The number of samples was to be increased as needed and would have been judged sufficient only if five independent repetitions of the analysis all yielded the same conclusions with regards to the parameter values and the model comparison. To define prior distributions, the model-fitting procedure outlined above was to be applied to the pilot data using broad priors (a normal distribution prior with the mean set to 0 and variance set to 100 for the α and β parameters; and an inverse gamma distribution with the shape and rate set to 0.01 for the σ^2 parameters). The resulting posterior distributions would serve as prior distributions for the main experiment data.

Model-fitting procedures were adapted to the exploratory modelling approaches (see 'Exploratory analyses').

Model comparison. To compare between pairs of models in terms of how well each model fit participants' biases, we planned to compute the evidence in favour 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.

Model comparison procedures were adapted to the exploratory modelling approaches (see 'Exploratory analyses').

Quality checks. To ensure that the collected data were able to test the study's hypothesis, we required three criteria. First, to ensure the quality of the pupil diameter data, we required that pupillary responses to oddball stimuli be significantly stronger than responses to the other stimuli in the auditory oddball task. Responses to each stimulus were computed as described above (see the section 'Eye tracking') and then averaged separately for oddball and non-oddball stimuli for each participant. A one-tailed paired t-test ($\alpha = 0.05$) across participants was used to determine whether responses to oddballs were indeed stronger. If this had not been the case, it would indicate that the pupillary recordings were not sufficiently sensitive even to capture this typically robust effect, or else that participants were not paying attention to the oddballs.

Second, since some of our inferences assumed a negative correlation between pupillary responses and baseline pupil diameter, we required that such anti-correlation be evident across participants in the pupil responses to oddball stimuli across the whole experiment. This anti-correlation was assessed by computing the Pearson correlation across trials between the oddball response and pre-stimulus baseline within each participant. We then conducted a one-tailed t-test across participants to determine whether the average correlation was indeed smaller than 0 ($\alpha = 0.05$).

Third, in the decision-making tasks, we required a statistically significant bias to be evident in at least one of the participant terciles when averaged across all six experimental tasks. To average biases (and pupil responses) across tasks, these were scaled such that 1 corresponded to the standard deviation across participants. Biases were then averaged for each participant and a one-tailed t-test across participants ($\alpha = 0.05$) was used to determine whether biases were indeed larger than zero in each of the participant groups. If biases had not been evident in

any of the groups, this would indicate a lack of replicability, or otherwise that our participant group might not have been sufficiently engaged in the experiment.

Exploratory analyses. To address feasibility issues that arose in the trial-level modelling, we complemented the planned analyses with two common trial-level modelling approaches that directly tested the study's hypothesis.

Hierarchical Bayesian modelling. We used STAN[39] in the R programming environment to infer the posterior distribution of the parameters of the trial-level model outlined above. To account for outliers, the model assumed that biases were drawn from a *t* distribution, although this was not formally tested. The normality parameter (ν) of the *t* distribution was drawn from an exponential prior with a mean of 30, as recommended in the literature[40]. The posterior distribution of this parameter (95% CI = 3.32 to 4.32) confirmed the presence of outliers. Predictors included the mean pupillary response of each participant (effect represented by β_3), the mean pupillary response of each participant for each task (orthogonalized with respect to the former predictor; β_2) and the trial-by-trial pupillary responses (orthogonalized with respect to both former predictors; β_1). To allow for changes across tasks in the effects of the pupillary responses, each β coefficient was computed as the sum of an average coefficient (drawn from a standard normal prior) and a task-specific coefficient (drawn from a normal distribution with a mean of 0 and a standard deviation that was a free parameter; subject to $\sum\beta = 0$ to regularize the model). Statistical significance was indicated by the absence of overlap between a parameter's 95% CI and a region of practical equivalence[40] interval, from −0.001 to 0.001, that was deemed practically equivalent to zero.

In addition to the pupillary response coefficients, a global intercept was drawn from the standard normal distribution, and question-specific ($\alpha_{t,q}$) and participant-specific (α_s) sets of intercepts (added to the global intercept and subject to the constraints $\sum\alpha_{t,q} = 0$ and $\sum\alpha_s = 0$) were each drawn from a normal distribution with a mean of 0 and a standard deviation that was a free parameter. Similarly, the global standard deviation of bias effects was drawn from a log-normal distribution with a mean of log[0.5] and a standard deviation of log[2], and question-specific ($\sigma_{t,q}$) and participant-specific (σ_s) sets of standard deviations (added to the global standard deviation in log space and subject to the constraints $\sum\log\sigma_{t,q} = 0$ and $\sum\log\sigma_s = 0$) were drawn from a log-normal distribution with a mean of 0 and a standard deviation that was a free parameter. The square of each top-level free standard deviation parameter was drawn from a weakly informative inverse gamma distribution whose two parameters were set to 0.01.

Three Monte Carlo Markov chains were sampled. Each chain's initial 500 samples were designated for warmup and discarded. A total of 10,000 samples were drawn subsequently. No divergent transitions were observed. For all parameters, the effective sample size was greater than 1,000 and $\hat{R}$ was lower than 1.1. Resulting posterior distributions are reported in terms of their mode, 95% credible interval and P_d.

Linear mixed modelling. Trial-by-trial data were also modelled using a linear mixed model, implemented with the lme4 package for R[41]. Here, rather than orthogonalizing predictors and examining the posterior distributions of the parameters, we used a model comparison approach whereby we eliminated terms that increased the BIC[38] or caused convergence problems or non-singular fits. The full model included fixed effects for the mean pupillary response of each participant (β_3), the mean pupillary response of each participant for each task (β_2) and the trial-by-trial pupillary responses (β_1). Random intercepts (α_s and $\alpha_{t,q}$) and slopes were included for participants and questions. The best-fitting model was determined by iteratively removing predictors until a minimal BIC was reached. This analysis assumed that data were normally distributed, but this was not formally tested.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study have been deposited on the Open Science Framework and are publicly available at https://osf.io/sygz3/.

Code availability

The custom scripts used for this study have been deposited on the Open Science Framework and are publicly available at https://osf.io/sygz3/.

Received: 29 July 2017; Accepted: 23 October 2020;
Published online: 04 January 2021

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. *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. Multialternative 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**, 757–769 (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**, 432–459 (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**, 370–392 (2001).
11. Johnson, J. G. & Busemeyer, J. R.A dynamic, stochastic, computational model of preference reversal phenomena. *Psychol. Rev.* **112**, 841–861 (2005).
12. Kahneman, D. *Attention and Effort* (Prentice-Hall, 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. Kahneman, D. *Thinking, Fast and Slow* (Macmillan, 2011).
22. Camerer, C. F. & Hogarth, R. M. The effects of financial incentives in experiments: a review and capital–labor–production framework. *J. Risk Uncertain.* **19**, 7–42 (1999).
23. De Gee, J. W., Knapen, T. & Donner, T. H. Decision-related pupil dilation reflects upcoming choice and individual bias. *Proc. Natl Acad. Sci. USA* **111**, E618–E625 (2014).
24. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).
25. De Gee, J. W. et al. Dynamic modulation of decision biases by brainstem arousal systems. *eLife* **6**, e23232 (2017).
26. De Gee, J. W. et al. Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains. *eLife* **9**, e54014 (2020).
27. Reas, C. & Fry, B. *Processing: a Programming Handbook for Visual Designers and Artists* (MIT Press, 2007).
28. Lambert, A., Wells, I. & Kean, M. Do isoluminant color changes capture attention? *Atten. Percept. Psychophys.* **65**, 495–507 (2003).
29. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
30. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
31. Jacowitz, K. E. & Kahneman, D. Measures of anchoring in estimation tasks. *Pers. Soc. Psychol. Bull.* **21**, 1161–1166 (1995).
32. Peterson, C. R. & DuCharme, W. M.A primacy effect in subjective probability revision. *J. Exp. Psychol.* **73**, 61–65 (1967).
33. 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).
34. 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).
35. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–292 (2013).

36. Shafir, E. Choosing versus rejecting: why some options are both better and worse than others. *Mem. Cogn.* **21**, 546–556 (1993).
37. Griffin, D. & Tversky, A. The weighing of evidence and the determinants of confidence. *Cogn. Psychol.* **24**, 411–435 (1992).
38. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
39. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
40. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan* (Academic Press, 2014).
41. Bates, D. et al. Package 'lme4'. *Convergence* **12**, 2 (2015).

Acknowledgements

This project was made possible through grants from the Israel Science Foundation (grant No. 1094/20; to E.E.), Army Research Office (grant No. W911NF-14-1-0101; to Y.N.) and Templeton Foundation (to V.F., J.D.C. and Y.N.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

E.E. conceived of the study. E.E., V.F., J.D.C. and Y.N. developed the methodology. E.E. and V.F. performed the investigations. E.E. and V.F. wrote the original draft of the manuscript. E.E., V.F., J.D.C. and Y.N. reviewed and edited the manuscript. Y.N. acquired funding. J.D.C. and Y.N. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-01006-3.

Correspondence and requests for materials should be addressed to E.E.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reprints and permissions information is available at www.nature.com/reprints.

Peer review information *Nature Human Behaviour* thanks Christopher Chambers, Joshua Gold, Peter Murphy and Konstantinos Tsetsos for their contribution to the peer review of this work. Primary Handling Editor: Marike Schiffer.

nature research

Corresponding author(s): Eran Eldar

Last updated by author(s): Sep 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data collection code has been deposited on the Open Science Framework and publicly available at: https://osf.io/sygz3/
Data analysis	All data analysis code has been deposited on the Open Science Framework and publicly available at: https://osf.io/sygz3/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data has been deposited on the Open Science Framework and publicly available at: https://osf.io/sygz3/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	quantitative experimental
Research sample	Participants were recruited from the greater Princeton area (mean age 19.9, age range 18-33; 105 female, 54 male). Inclusion criteria were age 18 to 35 and compatibility with pupillometry, as evidenced by successful calibration of the eye tracker. The sample was designed so as to minimize heterogeneity in order to avoid variance in decision making biases unrelated to pupil-linked mechanisms.
Sampling strategy	Participants self-selected by volunteering to the study in exchange for course credit or monetary compensation. Sample size was preregistered as determined by a power analysis conducted by bootstrapping pilot data.
Data collection	Data were collected using computer and eye tracker often with a non-blind experimenter in the room.
Timing	From 01-14-19 to 08-06-19 without substantial breaks.
Data exclusions	Exclusion criteria were predetermined. Data from 19 participants were not analyzed due to unsuccessful calibration of the eye tracker. Examination of the eye tracking data according to our a-priori exclusion criteria revealed that 8 participants had no valid data for any of the tasks, and 5 additional participants had no valid data for the oddball task. In addition, 5 participants were removed due to a programming error that prevented the tasks from running correctly, and 1 participant was removed due to having previously performed the experiment.
Non-participation	No subject declined participation
Randomization	Participants were not allocated into experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants self-selected by volunteering in exchange for course credit or monetary compensation. We do not expect this recruitment scheme influenced the results.
Ethics oversight	Princeton University Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.