Arousal, exploration and the locus coeruleus-norepinephrine system

Jepma, M.

Chapter 2

Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory

This chapter is published as: Jepma, M., & Nieuwenhuis, S. (in press). Pupil diameter predicts changes in the exploration-exploitation tradeoff: Evidence for the adaptive gain theory of locus coeruleus function. *Journal of Cognitive Neuroscience.*

Abstract

The adaptive regulation of the balance between exploitation and exploration is critical for the optimization of behavioral performance. Animal research and computational modeling have suggested that changes in exploitative vs. exploratory control state in response to changes in task utility are mediated by the neuromodulatory locus coeruleus-norepinephrine (LC-NE) system. Recent studies have suggested that utility-driven changes in control state correlate with pupil diameter, and that pupil diameter can be used as an indirect marker of LC activity. We measured participants' pupil diameter while they performed a gambling task with a gradually changing pay-off structure. Each choice in this task can be classified as exploitative or exploratory, using a computational model of reinforcement learning. We examined the relationship between pupil diameter, task utility and choice strategy (exploitation vs. exploration), and found that (i) exploratory choices were preceded by a larger baseline pupil diameter than exploitative choices; (ii) individual differences in baseline pupil diameter were predictive of an individual's tendency to explore; and (iii) changes in pupil diameter surrounding the transition between exploitative and exploratory choices correlated with changes in task utility. These findings provide novel evidence that pupil diameter correlates closely with control state, and are consistent with a role for the LC-NE system in the regulation of the exploration-exploitation trade-off in humans.

Introduction

Imagine you are in a restaurant, and are faced with the decision what food to order. One option is to choose a familiar dish that you know and like. Alternatively, you could try an unfamiliar dish, and take the risk that you might not like it. However, it is also possible that the unfamiliar dish turns out to become your new favorite, which you would never have discovered when sticking to the familiar dish. This example illustrates the dilemma between exploiting well-known options and exploring new ones. The trade-off between exploitation and exploration plays an important role in all kinds of decisions, especially in unfamiliar or changing environments. Although there has been a recent rise in studies investigating the strategies that are used to handle this trade-off and the neural mechanisms involved (for a review see Cohen, McClure, & Yu, 2007), these issues are still poorly understood.

One relevant line of research that addresses this issue suggests that the locus coeruleus-norepinephrine (LC-NE) neuromodulatory system plays an important role in regulating the balance between exploitation and exploration (Aston-Jones & Cohen, 2005; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). Aston-Jones and Cohen have proposed that exploitative and exploratory *control states* are mediated by two modes of LC activity, called the 'phasic' and the 'tonic mode', respectively. The phasic LC mode is characterized by an intermediate level of LC baseline activity and large phasic increases in activity in response to task-relevant stimuli. The ensuing phasic release of NE in cortical areas temporarily increases the responsivity (or gain) of these areas to their afferent input, selectively potentiating the processing of these task-relevant stimuli (Berridge & Waterhouse, 2003; Doya, 2002; Servan-Schreiber, Printz, & Cohen, 1990). Conversely, the tonic LC mode is characterized by an elevated level of LC baseline activity and tonic NE release, and the absence of phasic responses[1].

According to the adaptive gain theory (Aston-Jones & Cohen, 2005), the two LC modes promote exploitation and exploration by adaptively adjusting the responsivity of cortical neurons: the phasic mode produces selective increases in neuronal responsivity in response to task-related stimuli, thereby optimizing performance in the current task (i.e. exploitation). In contrast, the tonic mode produces a more enduring and less discriminative increase in neuronal responsivity. Although this degrades performance within the current task, it facilitates the disengagement of attention from this task and the processing of other non-task related stimuli and/or behaviors (i.e. exploration). A second assumption of the theory is that transitions between the phasic and tonic LC modes and corresponding control states are driven by online assessments of task-related utility carried out in ventral and medial frontal structures (Aston-Jones & Cohen, 2005). Consistent with this hypothesis, anatomical studies have shown that the primary neocortical projections to LC come from

[1] Whereas we discuss the phasic and tonic LC modes as distinct, they likely represent the extremes of a continuum of function. When we refer to the phasic or tonic LC mode, we mean a *more* phasic or tonic LC mode, not necessarily the extremes of the continuum.

orbitofrontal and anterior cingulate cortex (Aston-Jones et al., 2002; Rajkowski, Lu, Zhu, Cohen, & Aston-Jones, 2000; Zhu, Iba, Rajkowski, & Aston-Jones, 2004)—areas known to be responsive to task-related rewards and costs of performance (Botvinick, 2007; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). In order to adaptively regulate the balance between exploitation and exploration, utility assessments are integrated over both short (e.g., seconds) and longer (e.g., tens of seconds) timescales. If long-term utility is high, temporary decreases in utility augment the phasic LC mode, in order to restore task performance. Conversely, long-term decreases in utility drive the LC toward the tonic mode, which facilitates disengagement from the current task and exploration of alternative behaviors.

The adaptive gain theory has been supported mainly by computational modeling studies (Usher et al., 1999) and neurophysiological studies in monkeys that have used relatively simple tasks (Aston-Jones & Cohen, 2005). In contrast, with one notable exception (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010), there have been no tests of this theory in humans yet. In order to test the theory in humans, a non-invasive measure of LC activity is required. There is preliminary evidence that pupil diameter can provide such a measure: although it does not appear to be under direct control of the LC, pupil diameter is correlated with LC activity and thus may be useful as a "reporter variable" (Nieuwenhuis, de Geus, & Aston-Jones, in press). Rajkowski, Kubiak, and Aston-Jones (1993), for example, found a strong correlation in monkeys between baseline pupil diameter and tonic LC firing rate over the course of 90 minutes of performance in a target-detection task. Furthermore, a recent study that investigated how pupil diameter is related to experimental manipulations of task-related utility and behavioral indices of task (dis)engagement showed that pupil diameter varied in a way consistent with predicted LC dynamics (Gilzenrat et al., 2010). Specifically, this study showed that decreases in long-term utility and behavioral indices of task disengagement were associated with increased baseline pupil diameters and decreased pupil dilations, mirroring the high tonic and low phasic activity associated with the tonic LC mode. However, although this study assessed pupil effects associated with task (dis)engagement, it did not explicitly investigate the exploitation-exploration trade-off since participants were not given the opportunity to explore different task options.

Inspired by the recent evidence that pupil diameter might be used as an indirect index of LC activity, we measured participants' pupil diameter while they performed a 'four-armed bandit' task with a gradually changing pay-off structure in which the trade-off between exploitation and exploration is a central component (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Figure 1; Appendix). Optimal performance in this task requires a delicate balance between exploitative and exploratory choices. We examined whether the relationship between pupil diameter, control state and task-related utility was consistent with the two main assumptions of the adaptive gain theory, namely that LC mode regulates the trade-off between exploitative and exploratory control states, and that transitions between LC modes are driven by assessments of task-related utility. The first assumption predicts that exploratory choices will be associated with a larger baseline pupil diameter, possibly reflecting a more tonic LC mode, than exploitative choices. In addition, this assumption suggests that individual differences in overall pupil diameter might be correlated with individual differences in exploratory choice behavior: participants with larger overall pupil diameters, perhaps suggestive of a more tonic LC mode, may make more exploratory choices. The second assumption predicts that changes in utility surrounding the transition between control states will be accompanied by specific changes in baseline pupil diameter: a steady increase in baseline pupil diameter as decreasing utility drives the participant toward exploration; a monotonic decrease in baseline pupil diameter as utility increases after the participant has started a new series of exploitative choices.

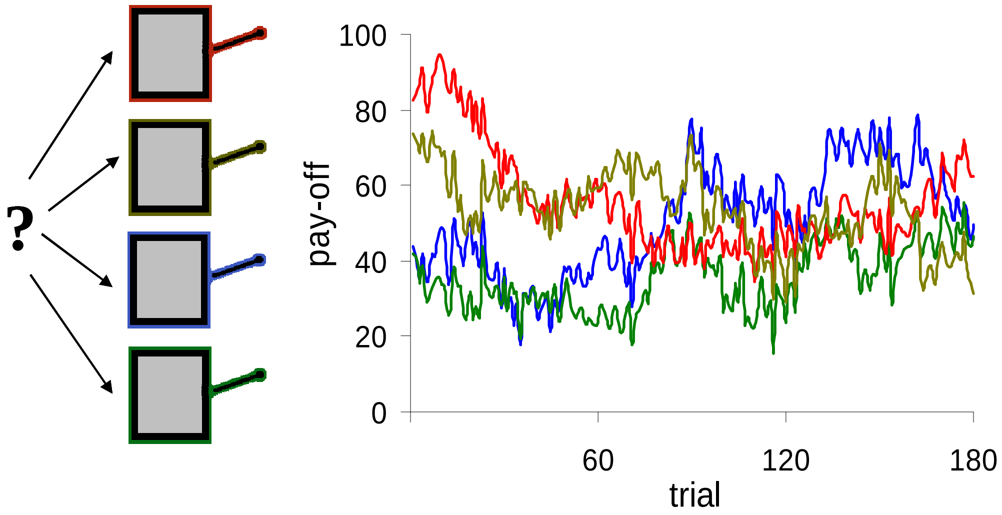

Figure 1. The four-armed bandit task. Participants made repeated choices between four slot machines. Unlike standard slots, the mean pay-offs of the four machines changed gradually and independently from trial to trial (four colored lines). Participants were encouraged to earn as many points as possible during the experiment. After the experiment, each choice was classified as exploitative or exploratory, using a computational model of reinforcement learning.

Materials and Methods

Participants

Seventeen volunteers participated (11 women; aged 18-33 years; mean age = 22.4). The experiment was approved by the local ethics review board and conducted according to the principles expressed in the Declaration of Helsinki. Informed consent was obtained from all participants.

Stimuli and Procedure

Participants performed a ‘four-armed bandit’ task, while their pupil diameters were continuously measured. The task was a slightly modified version of the task used by Daw et al. (2006). Participants were presented with pictures of four different colored slot machines (of equal luminance) on a medium gray background. The slot machines stayed on the screen during the entire experiment. Each trial started with a 4 s interval during which the slot machines were displayed, but participants could not select a machine yet. After this, a black fixation cross appeared in the center of the screen, indicating that participants could select one of the slot machines, by pressing the ‘q’-, ‘w’-, ‘a’- or ‘s’- key. Participants had a maximum of 1.5 s in which to make their choice; if no choice was made during that interval, a ‘TIME OUT’ message appeared in the center of the screen for 3 s to signal a missed trial (average number of missed trials = 1.7). If participants responded within 1.5 s, the lever of the chosen slot machine was lowered and the number of points earned was displayed in the chosen machine. These points were displayed until the end of the trial, which was 7 s after trial onset. Importantly, the number of points paid off by the four slot machines gradually and independently changed from trial to trial (Figure 1; Appendix).

The experiment was conducted at a slightly dimmed illumination level (room illumination 100 lux). We recorded pupil diameter at 60 Hz using a Tobii T120 eye tracker, which is integrated into a 17-inch TFT monitor (Tobii Technology, Stockholm, Sweden). Participants were seated at a distance of approximately 60 cm from the monitor. Prior to the start of the experimental session, participants viewed visually presented instructions, including an instruction that the pay-offs of the machines would change throughout the experiment, and were given 24 practice trials to familiarize them with the task. After the practice trials, participants were instructed that the machines had been reset for the experimental session. The experimental session consisted of 180 trials, and lasted about 20 minutes. We instructed the participants that they would be paid according to how many points they had earned during the experimental session. We also instructed them that on average participants earned 2.50 euros in this experiment. However, we did not tell participants how the number of points was converted into euros, or what their cumulative point total was. At the end of the experiment, each participant was paid 3 euros.

Data Analysis

Behavioral Analysis. In order to classify each choice as exploitative or exploratory, we fitted a reinforcement-learning model to the data of each participant. We used the same model as used by Daw et al. (2006). This model consists of a mean-tracking rule that estimates the mean pay-off of each machine, and a choice rule that selects a machine based on these estimations (Appendix). The choice rule we used was the ‘softmax’ rule. This rule assumes that choices between different options are made in a probabilistic manner, such that the probability that a particular machine is chosen depends on its relative estimated pay-off. The exploitation-exploration balance is adjusted by a parameter referred to as gain, or inverse temperature: with higher gain, action selection is determined more by the relative estimated pay-offs of the different options, whereas with lower gain, action-selection is more evenly distributed across the different options. We classified each choice as exploitative or exploratory according to whether the chosen slot machine was the one with the maximum estimated pay-off (exploitation) or not (exploration).

Pupil Analysis. Pupil data were processed and analyzed using Brain Vision Analyzer (Brain Products, Gilching, Germany). Artifacts and blinks were removed using a linear interpolation algorithm. We assessed the baseline pupil diameter prior to the selection of a slot machine, as well as the magnitude of the pupil dilation following the selection of a slot machine. To determine baseline pupil diameter, we averaged the pupil data in the period from 2.5 s to 0.5 s before the key-press. The pupil data during the 0.5 s immediately preceding the key-press were not included in the baseline period because most participants showed an anticipatory increase in pupil diameter starting about 0.5 s before their key-press response. The pupil dilation evoked by choosing a machine and perceiving the received pay-off was measured as the highest deviation from the baseline in the 3 s following the key-press response.

We compared the average baseline pupil diameter and pupil dilation on exploitation versus exploration trials. In addition, we calculated the degree of exploration for each exploratory choice, by subtracting the estimated pay-off of the chosen machine from the maximum estimated pay-off. We divided all exploration trials into three equally sized bins based on the degree of exploration (low, medium and high), and assessed the average baseline pupil diameter for these three exploration bins. Since the number of points earned was displayed immediately after the selection of a slot machine, the pupil dilation on each trial reflected both the selection of a machine and the processing of the received pay-off. Due to this confound, we could not unequivocally interpret differences in pupil dilation between exploitation and exploration trials, and focused our analyses on the baseline pupil diameter.

Compared to exploitative choices, exploratory choices were more often preceded by other exploratory choices. In addition, exploratory choices were associated with a lower pay-off and more negative prediction error on the previous trial, and a lower expected pay-off and higher entropy on the current trial (see Results). Entropy is an index of the similarity of the four slot machines' expected pay-offs; it increases as the expected pay-offs of the four slot machines become more similar. Entropy thus provides an estimate of the level of uncertainty, or conflict, associated with figuring out which slot machine is the most valuable. The entropy *H(X)* on each trial was calculated as:

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i)$$

where $P(x_i)$ is the probability of choosing slot machine x_i. To assess whether these potential confounds could account for the differences in baseline pupil diameter on exploration and exploitation trials, we subjected the single-trial baseline pupil diameter values to a multiple linear regression analysis, separately for each participant. Choice strategy (explore vs exploit) and the five above-mentioned nuisance variables (expected pay-off, entropy, and the pay-off, prediction error and strategy on the previous trial) as well as a constant were included as explanatory factors. For choice strategy and choice strategy on the previous trial, we used binary factors that have a value of 1 on exploit trials and 0 on explore trials. To assess which variables were significant predictors of baseline pupil diameter, we conducted a one-sample t-test on the regression coefficients of each explanatory factor (Lorch & Myers, 1990).

We also assessed whether individual differences in pupil diameter predicted individual differences in exploratory behavior. In this analysis, we computed the between-subjects correlation between the average baseline pupil diameter and the proportion of exploratory choices, and between the average baseline pupil diameter and the value of the gain/inverse temperature parameter of the reinforcement learning model.

To assess the development of our utility measures (pay-off, expected pay-off and entropy) and baseline pupil diameter surrounding the transition between exploitative and exploratory choice strategies, we averaged trials as a function of their position relative to the transition from an exploitative to an exploratory choice strategy, and vice versa. For this analysis, we only considered the exploration trials that were preceded or followed by a minimum of three exploitation trials.

Results

Participants alternated between choosing the slot machine with the highest estimated current pay-off (exploitation) and choosing slot machines with a lower expected pay-off (exploration). In comparison to the exploitation trials, exploration trials were more often preceded by other exploration trials (Table 1), indicating that participants tended to explore for several successive trials before settling on a new slot machine. The main characteristics of the exploitation and exploration trials are summarized in Table 1.

Table 1. Characteristics of exploration and exploitation trials (standard deviation in parentheses)

	exploration	exploitation	*p*-value
Proportion of total number of trials	0.31 (0.10)	0.69 (0.10)	< 0.001
Proportion preceded by exploration trial	0.41 (0.07)	0.28 (0.13)	0.001
RT (ms)	492 (82)	508 (75)	0.15
RT variability (SD of RTs)	150 (45.5)	151 (40.0)	0.912
RT trial N-1 (ms)	498 (72)	504 (79)	0.36
Pay-off (points)	48 (1.6)	63 (1.9)	< 0.001
Prediction error (points)	-2.8 (6.5)	-1.0 (5.1)	0.07
Expected pay-off (points)	51 (6.4)	64 (4.0)	< 0.001
Entropy (bits)	1.5 (0.14)	1.2 (0.33)	< 0.001
Pay-off preceding trial (points)	54 (2.4)	60 (3.1)	< 0.001
Prediction error preceding trial (points)	-3.6 (4.4)	-1.0 (5.9)	0.001

Pupil Diameter on Exploitation versus Exploration Trials

First, we compared the baseline pupil diameter preceding exploitative and exploratory choices. Baseline pupil diameters preceding exploratory choices were larger than those preceding exploitative choices [3.93 vs. 3.88 mm, $t(16) = 3.0$, $p = 0.008$; Figure 2, left panel]. Furthermore, within the exploration trials, baseline pupil diameter increased as a function of the degree of exploration (Materials and Methods), as revealed by a repeated-measures linear-trend analysis

[*F*(1,16) = 15.3, *p* = 0.001; Figure 2, right panel]. We also examined the pupil dilations evoked by exploratory and exploitative choices. There was a trend towards larger dilations on exploration than exploitation trials [0.17 vs. 0.13 mm; *t*(16) = 2.1, *p* = 0.051]. This was probably due to the higher incidence of negative prediction errors on exploration trials (Satterthwaite et al., 2007), since the effect disappeared when only the trials with positive prediction errors were included (*p* = 0.14).

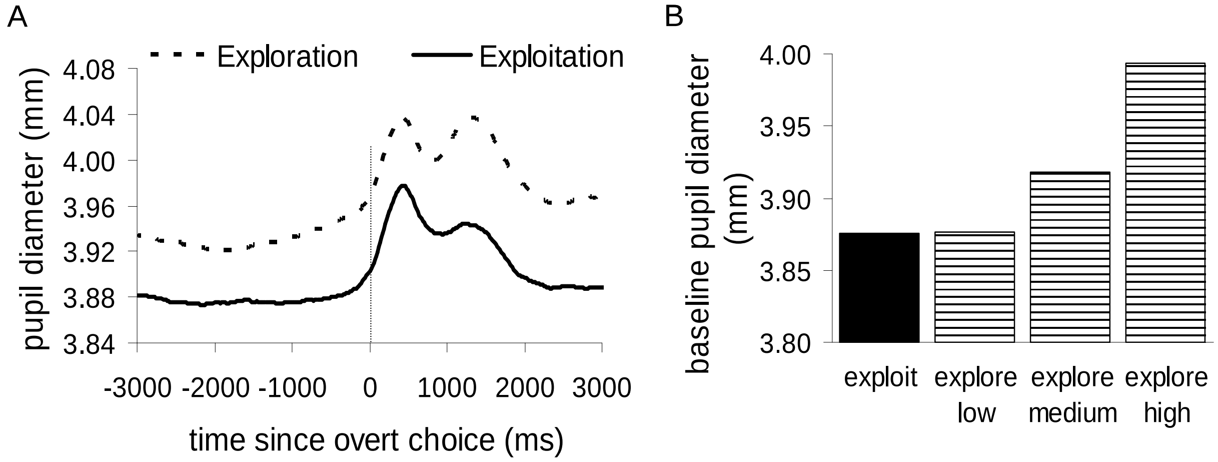

Figure 2. Pupil diameter on exploration and exploitation trials. (A) Time course of grand-average pupil diameter aligned to the key-pres indicating the selection of a slot machine, for exploratory and exploitative choices. (B) Average baseline pupil diameter for exploitative choices (black bar), and exploratory choices with a low, medium and high degree of exploration (striped bars).

The difference in baseline pupil size between exploitation and exploration trials already started to develop during the pupil response on the preceding trial (Figure 3): trials immediately preceding exploration trials were associated with a larger pupil dilation than trials immediately preceding exploitation trials [0.17 vs. 0.13 mm, *t*(16) = 3.2, *p* = 0.006]. However, this effect on the preceding trial could not (fully) explain the difference in baseline pupil diameters between exploitation and exploration trials, because the difference remained significant when pupil dilation on the previous trial was included as a covariate in the analysis [*F*(1, 15) = 4.69, *p* = 0.047].

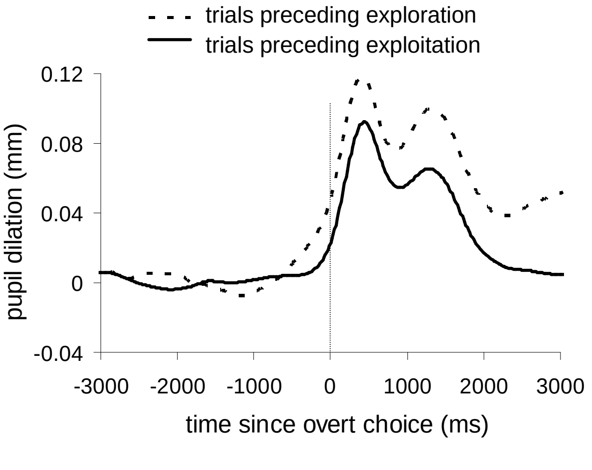

Figure 3. Time course of grand-average post-choice pupil dilation for the trials preceding exploration and exploitation trials.

Exploitation and exploration trials differed in several aspects other than choice strategy (Table 1). Trials preceding exploration trials were characterized by a larger proportion of exploratory choices, a lower pay-off and a more negative prediction error than trials preceding exploitation trials. In addition, exploration trials were characterized by a lower model-estimated expected pay-off (of the chosen slot machine) and higher entropy than exploitation trials. We investigated whether choice strategy (explore vs. exploit) could predict baseline pupil diameter independently of these potential nuisance variables by means of a linear multiple regression analysis (Materials and Methods). Importantly, when adjusted for all other variables, choice strategy made a unique contribution to the prediction of baseline pupil diameter [$t(16) = 3.43$, $p = 0.003$]. The only other significant predictor of baseline pupil diameter was the strategy on the previous trial [$t(16) = 2.98$, $p = 0.009$]. Additional control analyses that yielded similar results are reported in the Appendix.

Together, these findings confirm our first prediction that exploratory choices are associated with a larger baseline pupil diameter, while excluding a range of alternative interpretations for the observed pupil effect.

Individual Differences in Pupil Diameter and Exploratory Choice Behavior

Sofar we have examined pupil diameter as a function of the within-subject factor choice strategy. We next assessed whether individual differences in overall pupil diameter were predictive of individual differences in exploratory choice behavior. There was a positive correlation, across participants, between the average pupil diameter over all trials and the proportion of exploratory choices ($r = 0.50$, $p = 0.04$; Figure 4, left panel). Similarly, there was a negative correlation between the average pupil diameter and the value of the gain parameter of the reinforcement learning model ($r = -0.53$, $p = 0.03$; Figure 4, right panel). These correlations were also present when the baseline pupil diameters on exploitation and exploration trials were considered separately (pupil diameter on exploitation trials and proportion exploratory choices: $r = 0.49$, $p = 0.04$; pupil diameter on exploitation trials and gain parameter: $r = -0.52$, $p = 0.03$; pupil diameter on exploration trials and proportion exploratory choices: $r = 0.48$, $p = 0.05$; pupil diameter on exploration trials and gain parameter: $r = -0.53$, $p = 0.03$). Unlike the gain parameter, the other model parameters did not correlate with pupil diameter (decay parameter: $r = -0.24$, $p = 0.36$; decay center: $r = 0.07$, $p = 0.78$).

Obviously, individual differences in pupil diameter relate to many factors other than control state, such as age, personality and intelligence (Janisse, 1977). Importantly, these factors presumably increased the between-subjects error variance in our data, which decreased the power for detecting a correlation. Thus, the fact that we found a correlation in spite of a presumably large error variance in the between-subjects pupil data affirms the existence of the correlation. However, it is also possible that individual differences in pupil diameter reflect individual differences in motivation or the amount of attention paid to the task. Such motivational factors might influence choice strategy, which could provide an alternative explanation for the correlations between pupil diameter and exploratory behavior across participants.

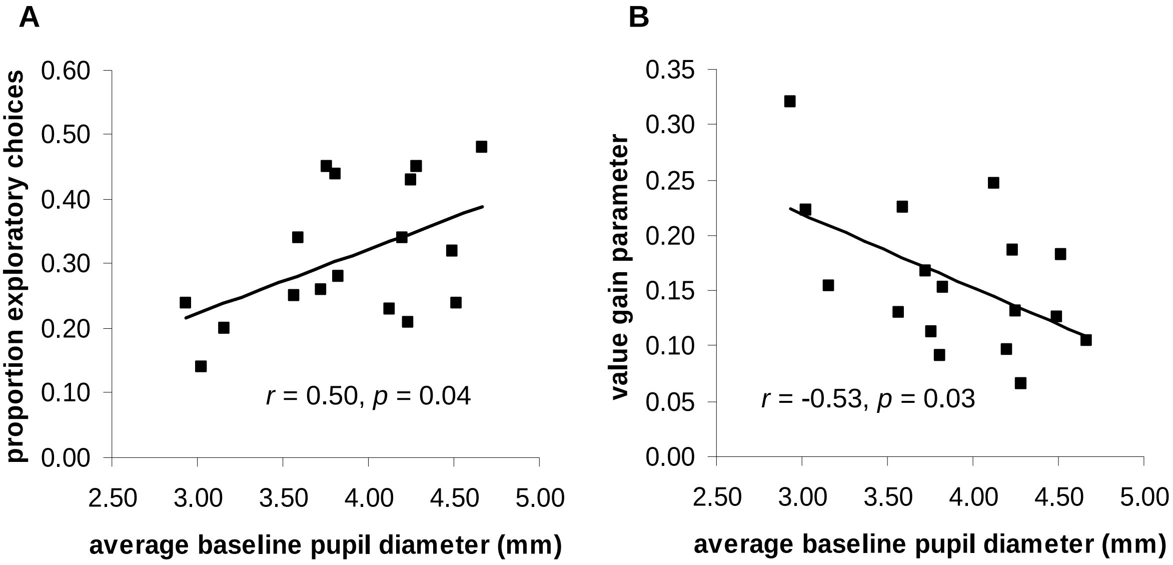

Figure 4. Individual differences in pupil diameter and exploratory choice behavior. (A) Scatter plot of the between-subjects correlation between average baseline pupil diameter and the proportion of exploratory choices. (B) Scatter plot of the between-subjects correlation between average baseline pupil diameter and the value of the gain or inverse-temperature parameter of the reinforcement-learning model. A lower value of this parameter indicates a more exploratory choice strategy.

Changes in Utility and Pupil Diameter Surrounding a Transition between Choice Strategies

Sofar we have examined the difference in pupil diameter between exploitation and exploration trials. We next examined the changes in utility measures surrounding the transition between exploitative and exploratory choice strategies. As measures of utility, we used the model-estimated expected pay-off of the chosen machine, the received pay-off, and the entropy (Materials and Methods). Subsequently, we tested whether such changes in utility were accompanied by changes in pupil diameter.

Figure 5 (upper panel) shows the expected pay-off, received pay-off and entropy for the first and the last of a series of exploration trials and the three preceding and following exploitation trials. During the three exploitation trials that preceded the switch to an exploratory choice strategy, entropy gradually increased [$F(1, 16) = 10.16$, $p = 0.006$] and pay-off gradually decreased [$F(1, 16) = 50.72$, $p < 0.001$], as revealed by a repeated-measures linear-trend analysis. Expected pay-off also showed a decrease over the three trials preceding the first explore trial, but this effect missed significance [$F(1, 16) = 2.85$, $p = 0.11$]. Thus, there was a gradual decrease in utility preceding the switch from an exploitative to an exploratory choice strategy, suggesting that, on average, participants began exploring when task utility was at a minimum. In addition, during the three exploitation trials following the last exploration trial, entropy gradually decreased [$F(1, 16) = 9.74$, $p = 0.007$] and expected pay-off gradually increased [$F(1, 16) = 13.72$, $p = 0.002$]. Thus, there was a gradual increase in utility following the switch from an exploratory to an exploitative choice strategy.

We next examined the development of baseline pupil diameter over the trials surrounding the switch between exploitative and exploratory choice strategies (Figure 5, lower panel). Baseline pupil diameter did not differ significantly across the three exploitation trials preceding the first exploration trial [$F(2, 32) = 1.30$, $p = 0.29$]. However, baseline pupil diameter showed a gradual decrease over the three exploitation trials following the last exploration trial [$F(1, 16) = 6.18$, $p = 0.024$], resembling the gradual decrease in entropy and increase in expected pay-off during these trials. As predicted, baseline pupil diameter correlated negatively with expected pay-off [$r = - 0.72$, p(1-tailed) = 0.023] and positively with entropy [$r = 0.68$, p(1-tailed) = 0.032] across the eight trial positions in Figure 5. These findings provide some evidence for our second prediction, that changes in utility surrounding the transition between control states would be systematically correlated with changes in baseline pupil diameter.

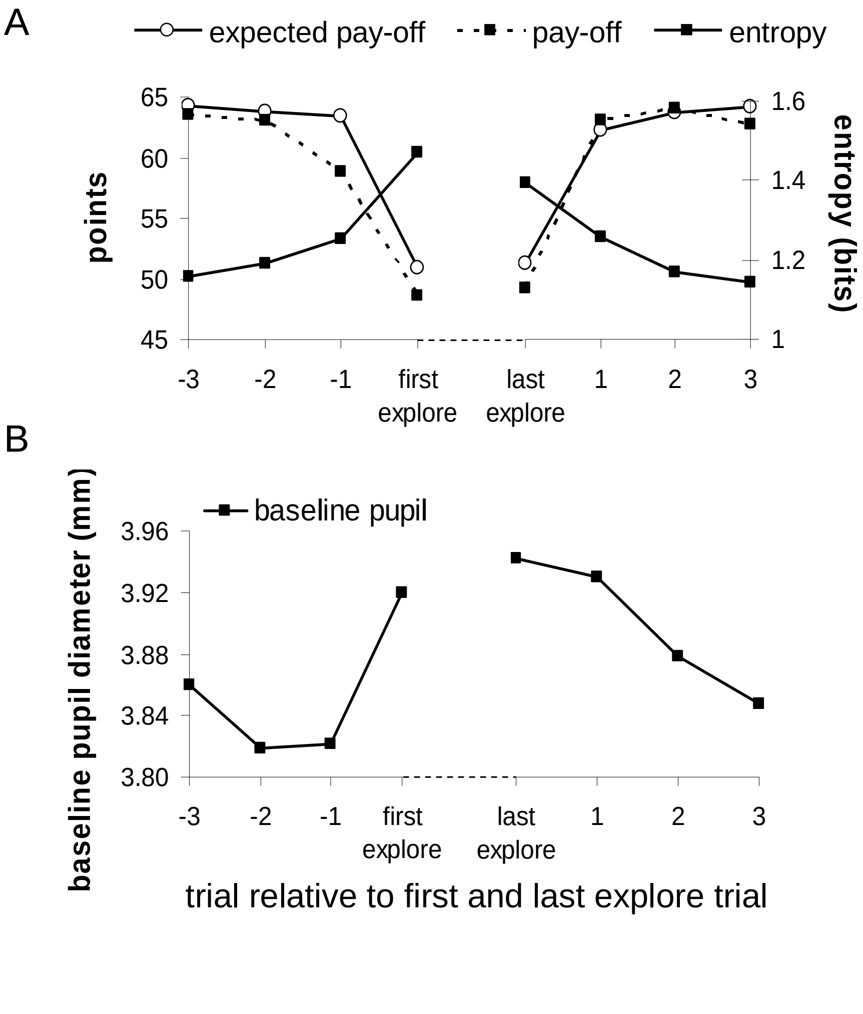

Figure 5. Grand-average dependent measures for the first and last of a series of exploration trials, and the three preceding and following exploitation trials. (A) Our measures of utility: expected pay-off, received pay-off and entropy. (B) Baseline pupil diameter.

Discussion

We investigated the relationship between pupil diameter, choice strategy (exploitation vs. exploration) and task utility, in order to test predictions of the adaptive gain theory of LC function in humans. This study was inspired by recent observations that pupil diameter might be used as a reliable index of LC activity. Our main findings can be summarized as follows: (i) exploratory choices were associated with a larger baseline pupil diameter than exploitative choices; (ii) individual differences in baseline pupil diameter predicted individual differences in exploratory choice behavior: participants with a larger pupil diameter made more exploratory choices and were characterized by a smaller gain parameter of the reinforcement-learning model; and (iii) trial-to-trial changes in baseline pupil diameter surrounding the transition between choice strategies correlated systematically with changes in utility, at least during the transition from exploration to exploitation. At the least, these findings provide novel evidence for a close relationship between pattern of pupillary response and control state. More tentatively, these findings provide indirect support for the two main assumptions of the adaptive gain theory, namely that LC firing mode regulates the trade-off between exploitative and exploratory control states, and that changes in LC mode are driven by online assessments of task-related utility (Aston-Jones & Cohen, 2005).

Our finding that pupil diameter is predictive of choice strategy, in a manner consistent with the adaptive gain theory, corroborates recent findings by Gilzenrat et al. (2010) that pupil diameter is related to behavioral indications of the tonic and phasic LC mode. Gilzenrat et al. found that large baseline pupils were associated with slower, more variable reaction times and less accurate performance in a target-detection task, and with task disengagement in a task in which participants were given the opportunity to disengage from the current task context when utility decreased. Furthermore, several pharmacological studies have shown that drug-induced activation of the LC-NE system increases cognitive flexibility and behavioral disengagement. For example, drugs that increase tonic NE levels (i.e. mimic the effects of elevated NE release that characterize the tonic LC mode) have been found to improve attentional-set shifting and reversal learning in rats and monkeys (Devauges & Sara, 1990; Lapiz & Morilak, 2006; Lapiz, Bondi, & Morilak, 2007; Seu, Lang, Rivera, & Jentsch, 2008; Steere & Arnsten, 1997; but see Chamberlain et al., 2006). In humans, increased NE levels induced by the selective NE reuptake inhibitor atomoxetine have been found to improve the ability to stop an ongoing motor response when cued to do so (Chamberlain et al., 2006). A possible explanation for this finding is that the drug-related increase in cognitive flexibility facilitates disengaging from one task (responding) and switching to a new task (stopping the response). In addition, increased NE levels induced by the selective NE reuptake inhibitor reboxetine have been found to enhance social flexibility in human participants, as indicated by increased social engagement and cooperation and a reduction in self-focus (Tse & Bond, 2002). Although none of these studies directly investigated exploitative versus exploratory behaviors, their findings support the idea that the tonic LC mode produces an enduring and largely nonspecific increase in responsivity, which promotes a flexible, exploratory control state.

Modeling studies have started to investigate the relationship between LC mode and task-related utility, integrated over different timescales (Aston-Jones & Cohen, 2005, Figure 10; McClure, Gilzenrat, & Cohen, 2005). However, to date there has been hardly any empirical research on the temporal dynamics of utility-driven changes in LC mode. We addressed this issue by assessing the trial-to-trial changes in utility and baseline pupil diameter surrounding the switch between exploitative and exploratory choice strategies. The switch to an exploratory choice strategy was preceded by a gradual decrease in utility, but an abrupt increase in baseline pupil diameter. When participants started to exploit a new machine after a period of exploration, utility gradually increased and baseline pupil diameter gradually decreased again. This pattern suggests that the transition from the tonic to the phasic mode is rather gradual, whereas the transition from the phasic to the tonic LC mode is more abrupt. A somewhat similar pattern was found by Gilzenrat and colleagues: Baseline pupil diameter showed a marked gradual decrease when participants started to engage in a new task; the increase in baseline pupil diameter leading up to task disengagement was less gradual and less pronounced. The implications of these data for our understanding of the specific mechanisms by which changes in short- and long-term utility control LC mode remain a matter for further research. One possibility is that LC baseline activity abruptly increases when long-term utility falls below a certain value. Consistent with this possibility, there is some evidence that tonic LC activity in monkeys can increase abruptly after a change in task contingency (Aston-Jones, Rajkowski, & Kubiak, 1997) or during the transition from a drowsy to an alert behavioral state (Rajkowski, Kubiak, & Aston-Jones, 1994). In any case, more empirical data is needed to determine how different measures of utility are integrated over different timescales and to specify the function relating overall utility to changes in LC mode. Such knowledge will also inform the implementation of a utility-sensitive adaptive gain mechanism in reinforcement-learning models. This will present a significant advance compared to current models, such as the model used here, in which the gain parameter is estimated for each participant but fixed across the experiment.

The abrupt increase in baseline pupil diameter prior to an exploratory choice might also be related to the specific task that we used. An aspect of the task that could be important in this respect is the high learning rate (see Appendix). A comparably high learning rate was found in a previous study using this task (Daw et al., 2006), so it seems to be characteristic of participants' choice behavior in this task. Such high learning rates imply that participants base their expectations regarding the slot machines' pay-offs primarily on their most recent experience with each machine. Accordingly, a single bad outcome on a certain trial is likely to be experienced as a substantial decrease in utility and to promote the exploration of another machine. This possibly explains the abrupt increase in baseline pupil diameter we observed immediately preceding the first of a series of exploratory choices. Thus, it will be important to assess in future studies whether tasks that are associated with lower learning rates will result in a more gradual increase in pupil diameter preceding the switch to an exploratory choice strategy.

Because the evidence for a close relationship between pupil diameter and LC activity is currently limited (Gilzenrat et al., 2010; Nieuwenhuis et al., in press; Rajkowski et al., 1993), more neurophysiological studies are needed to further establish this relationship. In addition, the neural mechanism underlying this putative relationship remains to be determined. To date, there are no known direct connections from the LC to the autonomic centers that regulate pupil size. It is more likely that pupil diameter and LC activity are closely linked because they receive downstream influences from a common afferent source. This common afferent might be the paragigantocellularis (PGi) nucleus of the ventral medulla, which plays a pivotal role in controlling both the LC and the sympathetic axis of the autonomic nervous system (Aston-Jones, Ennis, Pieribone, Nickell, & Shipley, 1986; Nieuwenhuis et al., in press). The notion that the LC and the autonomic nervous system receive their major input from a common source is consistent with several findings that suggest a strong temporal correlation between LC-NE activity and sympathetic nervous system activity (Elam, Svensson, & Thorén, 1986; Abercrombie & Jacobs, 1987; Reiner, 1986). Anatomical studies have revealed widespread afferents to the PGi from numerous brain areas, including the medial prefrontal cortex, insula, hypothalamus and periaqueductal grey, suggesting that activity in these areas might influence pupil diameter by way of the PGi (Aston-Jones et al., 1986). Consistent with this possibility, fMRI studies in humans and single-cell stimulation/recording studies in animals have shown that activity in this afferent network (including prefrontal cortex) is related to changes in pupil diameter (Critchley, Tang, Glaser, Butterworth, & Dolan, 2005; Loewenfeld, 1993; Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003).

Although our study focused on a possible role of the LC-NE system, it is unlikely that this is the only brain system involved in regulating the exploration-exploitation tradeoff. There is some evidence that the dopamine system also influences levels of exploration or task (dis)engagement (Dreisbach et al., 2005; Frank, Doll, Oas-Terpstra, & Moreno, 2009). For example, in one study, individuals with high spontaneous eyeblink rates (a marker of central dopaminergic activity) showed enhanced cognitive flexibility, as measured by the tendency to disengage from previously task-relevant stimuli and orient to novel stimuli (Dreisbach et al., 2005). Furthermore, this effect was modulated by the D4 dopamine receptor gene polymorphism. Another study reported that the val158met polymorphism of COMT, a gene that substantially affects prefrontal dopamine levels, could account for individual differences in uncertainty-based exploration (Frank et al., 2009). In addition to other neuromodulator systems, recent studies have implicated the frontopolar cortex in the control of exploratory behaviors (Daw et al., 2006; Bourdaud, Chavarriaga, Galan, & Millán, 2008), although the specific computations performed by the frontopolar cortex in this context are a topic of ongoing debate (Boorman, Behrens, Woolrich, & Rushworth, 2009). A key objective for future research is to specify the distinct contributions and interactions of the dopamine and LC-NE systems and the prefrontal cortex in the regulation of the exploration-exploitation tradeoff.

Our experimental design enabled examination of the baseline pupil diameter but, due to the overlap of the decision and outcome processing, did not allow examination of the decision-induced pupil dilation. Hence, the hypotheses we tested were restricted to the adaptive gain theory's assumptions about tonic LC activity. To provide complementary data with regard to phasic LC responses, an important aim for future studies is to use a task in which the decision and outcome presentation are separated in time such that the pupil dilations associated with these two processes can be isolated.

The present study tested specific predictions regarding the relationship between pupil diameter, utility measures and choice strategy based on a mechanistic theory about the role of the LC-NE system in regulating control state, and preliminary evidence from previous studies for a close relationship between LC activity and pupil diameter. Given the specificity, and therefore the intrinsic unlikelihood, of our predictions, the fact that the predicted effects were observed lends provisional support to the hypotheses that drove the predictions. However, since this is an inductive argument, it is important to note that we cannot rule out the possibility that the observed relationships were not related to LC-mediated modulation of control state. Thus, future studies using more direct measures or manipulations of the LC-NE system are needed to either confirm or invalidate the conclusions from the present study.

For a long time, the LC-NE system has been associated with basic functions such as arousal and the sleep-wake cycle. Only recently, researchers have begun to examine its involvement in more specific cognitive functions, such as attention, memory, perceptual selection and the signaling of unexpected uncertainty (Einhäuser, Stout, Koch, & Carter, 2008; Robbins, 1997; Sara, 2009; Yu and Dayan, 2005). The present study contributes to this work by addressing, albeit indirectly, the role of the LC-NE system in the control of human behavior. Specifically, the findings reported here support the adaptive gain theory (Aston-Jones & Cohen, 2005), which posits an important role for the LC-NE system in the optimization of behavioral performance by regulating the balance between exploitative and exploratory control states.

Appendix

Pay-off structure of the gambling task

The number of points paid off by slot machine *i* on trial *t* ranged from 1 to 100, drawn from a Gaussian distribution (standard deviation σ_o = 4) around a mean $\mu_{i,t}$ and rounded to the nearest integer. On each trial, the means diffused in a decaying Gaussian random walk:

$$\mu_{i,t+1} = \lambda\mu_{i,t} + (1-\lambda)\theta + \nu \; .$$

The decay parameter λ was 0.9836, the decay center θ was 50, and the diffusion noise ν was zero-mean Gaussian (standard deviation σ_d = 2.8). We used one instantiation of this process (Figure 1).

Reinforcement-learning model

We used a Bayesian mean-tracking rule (i.e. a Kalman filter) that tracked the mean expected pay-off of each machine $(\hat{\mu}_{i,t})$ and the variance of these pay-offs $(\hat{\sigma}^2_{i,t})$. On the first trial of the task, all four machines had the same prior mean $\hat{\mu}^{pre}_{i,1}$ and variance $\hat{\sigma}^{2\,pre}_{i,1}$. These start values were based on the pay-offs received during the practice block, and were determined separately for each participant (mean $\hat{\mu}^{pre}_{i,1}$ = 51.9, SD = 2.7; mean $\hat{\sigma}^{2\,pre}_{i,1}$ = 52.3, SD = 14.9). When a participant chose machine *c* on trial *t* and received pay-off *r*, the estimated pay-off distribution ($\hat{\mu}^{post}_{c,t}, \hat{\sigma}^{2\,post}_{c,t}$) was updated according to:

$$\hat{\mu}^{post}_{c,t} = \hat{\mu}^{pre}_{c,t} + \kappa_t \delta_t$$

$$\hat{\sigma}^{2\,post}_{c,t} = (1-\kappa_t)\hat{\sigma}^{2\,pre}_{c,t}$$

with prediction error $\delta_t = r_t - \hat{\mu}^{pre}_{c,t}$ and learning rate $\kappa_t = \hat{\sigma}^{2\,pre}_{c,t} / (\hat{\sigma}^{2\,pre}_{c,t} + \hat{\sigma}^2_o)$.

The estimated pay-off distributions for the unchosen machines did not change.

Then, the estimated prior pay-off distributions on the subsequent trial (trial *t*+1) were updated in time according to:

$$\hat{\mu}^{pre}_{i,t+1} = \hat{\lambda}\hat{\mu}^{post}_{i,t} + (1-\hat{\lambda})\hat{\theta}$$

$$\hat{\sigma}^{2\,pre}_{i,t+1} = \hat{\lambda}^2 \hat{\sigma}^{2\,post}_{i,t} + \hat{\sigma}^2_d \; .$$

We modeled the choice of the participants by a softmax rule. The probability $P_{i,t}$ of choosing machine *i* on trial *t* was given by:

$$P_{i,t} = \frac{\exp(\beta\hat{\mu}^{pre}_{i,t})}{\sum_j \exp(\beta\hat{\mu}^{pre}_{j,t})}$$

with exploration parameter β (often referred to as gain, or inverse temperature).

For a discussion of the Kalman filter and the softmax rule, we refer the reader to Anderson and Moore (1979), and Sutton and Barto (1998), respectively.

We fitted the model to each individual participant's choice data. The trials in which no response was made within the 1.5-s time limit were omitted. The parameters $\hat{\lambda}$, $\hat{\theta}$ and β were estimated per participant by maximizing the log-likelihood of the observed choices (Supplemental Table 1). Parameter $\hat{\sigma}_o$ was fixed at 4. Estimation of parameter $\hat{\sigma}_d$ resulted in extreme values for most of the participants (values larger than 1000 for ten of the seventeen participants), suggesting unreliable fits. Therefore, we fixed this parameter at 50, which is similar to the best fitting $\hat{\sigma}_d$ parameter found in a previous study (Daw, O'Doherty, Dayan, Seymour, and Dolan, 2006). This large value of $\hat{\sigma}_d$ implies that participants overestimate the speed of diffusion in the pay-offs. Large values of $\hat{\sigma}_d$ induce high learning rates, indicating that the expected pay-offs are determined primarily by the most recent experience with each machine.

Supplemental Table 1. Mean parameter estimates and negative log likelihood for the fit of the softmax model to the choice data of each participant. The parameter values used to generate the pay-offs, and the negative log likelihood of a model in which choices are made randomly are also shown.

	Estimated values	Generative values
β	0.160 (0.066)	
λ	0.894 (0.083)	0.9836
θ	56.9 (17.6)	50
σ_d	50 (fixed)	2.8
σ_0	4 (fixed)	4
-LL	153.1 (34.8)	
-LL randomly choosing model	247.2 (2.0)	

Note: SD in parentheses; -LL = negative log likelihood

Additional control analysis

Besides the multiple regression analyses, we performed a second set of control analyses to investigate whether differences in each of the potential confound variables could account for the different baseline pupil diameter on exploration and exploitation trials (and hence might provide an alternative interpretation of the effect). We repeated the comparison of baseline pupil diameter on exploitation and exploration trials while, in separate analyses, controlling for differences in each of the potential confound variables (pay-off on the previous trial, prediction error on the previous trial, expected pay-off on the current trial and entropy on the current trial), by matching the values of these variables across exploration and exploitation trials (Bernstein, Scheffers, & Coles, 1995). We sorted each participant's exploitation and exploration trials by one of these variables, and then successively removed the most extreme exploitation and exploration trials, thereby reducing the difference between the mean value of this confound variable on exploitation and exploration trials. After each trial removal, we calculated the difference between the mean values of the confound variable on exploitation and exploration trials, and we stopped the removal process when this difference was not further decreased by removal of a subsequent trial (Supplemental Table 2). We also controlled for choice strategy on the previous trial, by including only the trials that were preceded by an exploitation trial. Finally, in order to control more explicitly for the possibility that the higher incidence of negative prediction errors preceding exploratory choices was driving the effect, we repeated the analysis while only including the trials that were preceded by a positive prediction error.[2]

Importantly, none of the potential confound variables could account for the larger baseline pupils preceding exploratory compared to exploitative choices: the critical effect remained significant after correction for choice strategy on the previous trial [$t(16) = 2.5$, $p = 0.026$]; pay-off on the previous trial [$t(16) = 2.9$, $p = 0.009$]; prediction error on the previous trial [$t(16) = 2.5$, $p = 0.025$]; expected pay-off [$t(12) = 3.1$, $p = 0.010$]; and entropy [$t(16) = 3.5$, $p = 0.003$]. Furthermore, the effect remained significant when only the trials that were preceded by a positive prediction error were considered ($t(12) = 2.3$, $p = 0.037$), suggesting that the larger baseline pupil on explore compared to exploit trials was not due to the larger incidence of negative prediction errors preceding explore trials.

Supplemental Table 2. The number of excluded trials and the values of the potential confound variables on exploration and exploitation trials after correction.

	# excluded trials	Exploration	Exploitation	*p*-value
Expected pay-off	83.8 (13.5)	60.0 (4.1)	60.1 (4.0)	.29
Entropy	27.0 (14.0)	1.25 (0.30)	1.26 (0.29)	.02
Pay-off preceding trial	23.2 (15.9)	57.8 (2.0)	57.9 (2.0)	.13
Prediction error preceding trial	14.3 (9.6)	-1.81 (5.24)	-1.83 (5.28)	.40

Note: SD in parentheses. The difference in entropy after correction is in the opposite direction (larger entropy on exploitation trials) compared to the original effect.

Uncertainty-driven exploration and pupil diameter

In the softmax rule described above, the probability that a particular machine is chosen is determined by its relative mean estimated pay-off (and the value of the gain parameter), but not by the uncertainty about its potential pay-offs (i.e. the variance of the estimated pay-off distribution $\hat{\sigma}_i^{2\,pre}$). On the other hand, modeling studies have suggested that exploration might be directed towards particular choices in proportion to the uncertainty about their outcomes, which can be implemented by adding an ‘uncertainty bonus’ to the expected value of options with uncertain outcomes (e.g., Sutton, 1990). It has recently been shown that individual differences in uncertainty-

[2] Four participants had to be excluded from the analysis that corrected for expected pay-off, because the difference in expected pay-off between their exploration and exploitation trials was so large that no exploration trials were left using this procedure. Similarly, four participants were excluded from the analysis in which only the trials preceded by a positive prediction error were considered, since less than ten explore and/or exploit trials were left for these participants.

based exploration are associated with the val158met polymorphism of the COMT gene, which substantially affects prefrontal dopamine levels (Frank, Doll, Oas-Terpstra, & Moreno, 2009). According to the adaptive gain theory, the increased NE level in the tonic LC mode indiscriminately facilitates processing of all stimuli and/or behaviors, which promotes a nonspecific type of exploration. Hence, the theory predicts that individual differences in tonic LC activity (as indexed by baseline pupil diameter in this study) will be related to individual differences in exploratory behavior (Results section), but not to individual differences in uncertainty-specific exploration.

To asses this last prediction, we considered a softmax rule in which an ‘uncertainty bonus’ of φ standard deviations was added to the estimated mean pay-offs:

$$P_{i,t} = \frac{\exp(\beta[\hat{\mu}_{i,t}^{pre} + \varphi\hat{\sigma}_{i,t}^{pre}])}{\sum_j \exp(\beta[\hat{\mu}_{j,t}^{pre} + \varphi\hat{\sigma}_{j,t}^{pre}])}$$

The best fitting uncertainty bonus parameter in this model varied across participants: four participants had a positive bonus and thirteen participants had a negative bonus (mean bonus = -0.117, SD = 0.336). Thus, for the majority of the participants, uncertainty about the potential outcomes of a machine *discouraged* exploration of that machine. Importantly, the value of the uncertainty bonus parameter did not correlate with baseline pupil diameter (*r* = 0.05, *p* = 0.86), consistent with the assumption that the tonic LC mode is not associated with uncertainty-specific exploration.