Supporting Information

Hämmerer et al. 10.1073/pnas.1712268115

SI Methods

Experimental Procedure. Stimuli were presented and responses were recorded using the Cogent 2000 toolbox (www.vislab.ucl.ac.uk/cogent.php) running on MATLAB (version 2015a; MathWorks, 2015). Stimuli as well as background were luminance-controlled to avoid confounds in pupillometric recordings. Responses were recorded with scanner-compatible button boxes (fiber optic response pad HHSC-1X4-CL; Current Designs). The task included 303 trials of 6- to 18-s duration (mean duration 10 s), and was divided into six blocks of 8.5 min each. Note that five younger adults followed a slightly different temporal schedule (due to an added pilot sample). They completed 330 trials over five blocks of 66 trials each. Importantly, although trial numbers were slightly different, participants did not differ on the relative proportion of gains or losses received [$t(1,48) = 1.6$, $P = 0.12$].

Behavioral Analyses. Mean accuracy on the reversal learning task was assessed as the percentage of correct responses on free choice trials. Memory performance on recognition tests was assessed as hits–false alarms. In the assessment of memory performance, all trials that were responded to were included and analyzed according to the feedback valence on the trial. The percentage of hits or misses for subcategories of the data (e.g., scenes before losses) was determined as the number of correct old responses or incorrect new responses on this category compared with the number of old scenes for this category (e.g., overall number of old scenes before losses during recognition tests).

Multiple (logistic) regression analyses were used to explore whether a loss or a gain that occurred on up to three trials before or after the incidentally encoded scene stimulus was related to encoding success on the current trial. Specifically, the regressor coding for outcomes (0 for gains; 1 for losses) and predicting memory success (1 for hit; 0 for miss) was shifted up to three trials before and after the scene stimulus on a current trial (n) to examine whether outcome type on preceding or following trials had an impact on the memory for the scene stimuli of the current trial (n). Type of scene stimulus (indoor or outdoor), type of trial (free choice or forced), and reversal (1) versus no reversal trial (0) were included as regressors of no interest to control for different trial types when investigating the effects of outcomes on memory. Furthermore, to account for a possible autocorrelation among outcome regressors across continuous trials, outcomes on the three preceding as well as three following trials were also included as regressors of no interest. Correlations between regressors were on average below $r = 0.20$ (average across absolute values of r), suggesting that regressions were not affected by multicollinearity of regressors. Permutation tests were used to determine trial positions (e.g., $n = 1$ or $n = -3$), where outcome type significantly predicted hits (comparison against time series across randomly shuffled hits and misses per permutation based on 100 repetitions). Regression analyses which only focused on forced trials used all data but set the feedback on choice trials to the mean of the feedback across all trials to exclude its relevance on the regression results.

Pupillometric Recordings and Analyses. All pupillometric data were concurrently recorded during brain image acquisition in the reversal-learning task using a scanner-compatible eyetracking camera mounted behind the scanner bore (EyeLink 1000 Plus; SR Research, 2010). Changes in pupil diameter were measured from the right eye with a sampling rate of 1,000 Hz. The eye-tracking camera and infrared light projection were calibrated at the beginning of the first session and if necessary before the start of each task block with respect to optimizing pupil capture as well as maximizing infrared light intensity. Participants were asked to minimize blinking during feedback presentation.

Pupillometric analyses followed standard procedures as described in ref. 1. Pupil data were segmented 500 ms before and 2,500 ms after feedback presentation. Then, missing data due to eye blinks were linearly interpolated using in-house written scripts in MATLAB in time windows of 30 and 200 ms around small and large blinks, respectively. Next, each trial was individually inspected for artifacts using FieldTrip [version 20140615 (2)] and excluded if recordings were too noisy or too large quantities of data per trial had to be interpolated due to blinking. Pupil responses were then z-scored across all remaining trials per individual to compare pupillary response differences across trial types independent of interindividual differences in pupil size. Pupil responses were not baseline-corrected to assess both tonic as well as phasic components of pupil diameters to obtain a measure of how much functional activation of the LC should be overall expected during feedback. To compare pupil responses across conditions and participants, we used the mean in a time window of length 1 s, starting 1 s after feedback onset as well as across gain, loss, or reversal trials for every individual.

Due to the more challenging pupil recordings in long-range mount, a total of six participants (of which five were older adults) had to be excluded as they did not have sufficient trial numbers for analyses. There were no age-related differences in mean trial numbers per condition after artifact correction (mean trial numbers for gain trials and loss trials in younger adults, 67.95 and 26.40, and in older adults, 78.82 and 28.64).

Structural MRI Analyses. Individual scans were inspected for movement artifacts during recording and if necessary repeated. For LC mask definition, the T1-weighted multiecho FLASH scans were averaged across six repetitions as well as six echo times to increase SNR. Age groups did not differ in variance of movement regressors across the six repetitions ($t = -1.02$, $P = 0.32$). Also, variance of movement regressors did not contribute significantly to NM signal intensity as determined by ratio score (see below) within the LC mask or size of the LC mask (younger adults: $r = 0.18$, $P = 0.44$ and $r = 0.18$, $P = 0.44$, respectively; older adults: $r = -0.25$, $P = 0.19$ and $r = 0.14$, $P = 0.49$, respectively). This suggests that interindividual differences in movement between acquisitions did not affect our assessment of LC integrity.

On T1-weighted images, the LC appears as voxels of high intensity in the lateral floor of the fourth ventricle (Fig. S4 for individual T1-weighted images). The LC was initially defined on the $0.4 \times 0.4 \times 3.0$-mm anisotropic T1-weighted multiparametric mapping scans using ITK-snap (3). Then, to further improve the localization of the superior and inferior boundaries, data were coregistered to the 0.75-mm^3 isotropic T1-weighted scans using SPM12 (Statistical Parametric Mapping, www.fil.ion.ucl.ac.uk/spm), and information from both images was used to refine the original segmentations in ITK-snap. The LC mask was defined as the conjunction of labeled voxels from two raters. Interrater reliability was assessed as the correlation in numbers of voxels selected per person as well as Dice similarity coefficient (DSC) score [DSC(A, B) = 2 (conj(A_B))/(A + B)] (4); interrater reliability number of voxels: $r = 0.78$, $P < 0.05$; mean % of same voxels: 57%, no age-related difference (younger adults, 60%; older adults, 54% [$t(1,48) = -1.65$, $P = 0.10$]; DSC = 0.72, no age-related difference, younger

adults = 0.74, older adults = 0.70 [$t(1,48) = 1.48$ $P = 0.15$]). Masks were then coregistered back to the anisotropic space of the T1-weighted images. Coregistration success was visually assessed. To improve coregistration performance, the mean T1-weighted images were bias-corrected before coregistration using SPM12. To generate a metric for comparison across individuals, a reference region of interest was drawn in the dorsal pons close to the LC masks. Given the relatively flat receive fields for central regions, a control area close to the LC should be expected to be less affected by differences in receive sensitivity. The control area was 10 × 10 voxel (that is, 4 × 4 mm)-wide and situated on the middle slice of the LC mask in the z direction, halfway between left and right LC masks in x (left–right) direction and 5 voxels (2 mm) above LC masks in the y direction (rostro–caudal) (Fig. S5). The ratio score of signal intensity (SI) in LC masks was then calculated as the mean across voxels for (LCSI − pons controlSI)/pons controlSI (5).

Statistical Procedures. t tests were used to assess age group differences in the behavioral, pupillometric, or structural MRI data. Age group differences in recognition performance were further assessed using repeated measures ANOVAs, with age groups as fixed effect and recognition performance across test time points or recognition performance across scenes before loss or gain feedback as repeated measures. Condition as well as age group differences in correlations of behavioral and structural MRI data were assessed using permutation tests. Analyses were carried out using SPSS version 24.0.0.1 (IBM; https://www.ibm.com/analytics/de/de/technology/spss) and MATLAB version R2016b (The MathWorks; 0.1.0.441655). Permutation tests on correlations and regression analyses were carried out in MATLAB.

SI Results

Control Analyses on Difference in False Alarms Between First and Second Memory Tests. The correlation of NM signal intensity and memory for stimuli before losses in older adults was not substantially affected by including the difference in false alarms in a partial correlation. In line with this, while there was a negative association for the difference in false alarms with NM signal intensity in older adults ($r = -0.43$, $P < 0.05$), it was not reliably associated with the memory measure ($r = -0.29$, $P = 0.19$), thereby explaining the lack of a substantial reduction when included in the partial correlation above. Rather, it seems that NM signal intensity appears to share additional variance with the difference in false alarms, such that older adults with lower NM signal intensity show comparatively more false alarms on the first test (correlation NM signal intensity and false alarms on first test: $r = 0.44$, $P < 0.05$; second test: $r = -0.32$, $P = 0.14$).

Control Analyses on Impact of Voxels Included in Mask on NM Signal Intensity Measure. We did not observe a reliable correlation between the number of voxels included in the mask and mean signal intensity across all voxels in older adults ($r = 0.02$, $P = 0.93$). Also, the median of signal intensities was highly correlated to the mean signal intensity ($r = 0.95$, $P < 0.05$) and reliably related to the relevant memory measure ($r = 0.47$, $P < 0.05$). This suggests that the number of voxels included in the mask did not substantially affect mean signal intensities via a dilution by more less-intense voxels in larger masks.

Memory Performance: Familiarity or Recollection. In addition to indicating whether a scene stimulus was new or old, participants were asked to indicate whether they remembered (that is, recalled aspects of the memory episode explicitly) or knew (that is, had a sense of familiarity about the stimulus without explicit recollection) that the scene was old. To differentiate the effect of positive or negative feedback on familiarity or recollection of the scene stimuli, we therefore examined scenes before gains or losses that were remembered or judged familiar (repeated measures ANOVA gain versus loss feedback x remember versus know scene before feedback x age group). As in the overall memory analyses, we observed a main valence effect of remembering or knowing scene stimuli that were presented before losses compared with stimuli presented before gains [$F(1,48) = 12.69$, $P < 0.05$, $rICC = 0.46$]. In addition, we observed a main effect of memory type. Participants indicated more often to know rather than remember a scene [$F(1,48) = 13.86$, $P < 0.05$, $rICC = 0.47$], suggesting that familiarity with the presented stimuli outweighed recollection of the presented stimuli. We also observed a trend for an age group x memory type interaction indicating older adults endorsed more frequently know rather than remember in response to an old stimulus [$F(1,48) = 3.96$, $P = 0.05$, $rICC = 0.28$]. There was no interaction with the valence of the feedback.

Regarding correlations with the NM signal intensity measure, better memory performance with higher NM signal intensity was only observed in older adults for know responses for stimuli before losses (older adults: $r = 0.50$, $P < 0.05$; younger adults: $r = 0.25$, $P = 0.19$), although a trend was also observed for know responses for stimuli before gains (older adults: $r = 0.38$, $P = 0.08$; younger adults: $r = 0.30$, $P = 0.12$), whereas remember judgments were not reliably correlated with NM signal before losses (older adults: $r = 0.26$, $P = 0.25$; younger adults: $r = -0.16$, $P = 0.56$) or before gains (older adults: $r = -0.02$, $P = 0.91$; younger adults: $r = -0.21$, $P = 0.29$). Note that the median of the number of correctly remembered hits that could be entered in the overall analysis (integrated across gain and loss feedback trials) was quite low—only 23 stimuli per participant—whereas correctly known hits were comparatively more frequent (median of 60). Remember memory results might thus have to be interpreted with more caution due to insufficient trial numbers to obtain reliable estimates of individual differences in recollection.

Forced Trials Only: Memory on Trials Preceding or Following Loss Versus Gain Outcomes. Regression analyses on forced trials (Fig. S1) only confirmed the pattern of results observed on all trials (correlation between NM signal intensity and memory for stimuli before losses: older adults: $r = 0.68$, $P < 0.05$; younger adults: $r = 0.05$, $P = 0.83$; correlation between NM signal intensity and memory for stimuli before gains: older adults: $r = 0.10$, $P = 0.65$; younger adults: $r = 0.05$, $P = 0.83$; correlation between NM signal intensity and memory for all forced stimuli: older adults: $r = 0.48$, $P < 0.05$; younger adults: $r = 0.11$, $P = 0.55$).

Pupil Responses to Reversal Trials. Younger adults as well as older adults responded with a strong phasic change in pupil diameter (PD) to outcomes indicating a reversal trial compared with nonreversal trials [repeated measures ANOVA, fixed factor age group x gain, loss, and reversal trials $F(1,76) = 6.88$, $P < 0.05$, $rICC = 0.29$; no age interaction]. Across both age groups, responses to reversals were larger than responses to gains [$t(1,40) = 3.46$, $P < 0.05$] but not reliably larger than responses to losses [$t(1,39) = 1.42$, $P = 0.16$]. Although older adult responses to reversals seem substantial, we did not observe an age interaction in PD responses to reversal, loss, and gain trials. It is possible, however, that with a larger sample, the age interaction would have been reliable. There are currently very few studies which compare PD effects during cognitive tasks in younger and older adults, and larger pupil diameters in older adults might seem surprising given the assumed overall lower levels of noradrenergic modulation (of which PD is conceived to be a proxy measure) in older adults. However, the size of PD will not only depend on the overall potency of the noradrenergic system but also on the strength of the attentional focus based on task sets within which a salient event is evaluated (6, 7). An increased task state focus (especially on lower levels of task complexity) in older adults is not unprecedented (8),

and might explain increased PD effects in older adults especially when establishing an attentional focus is easier. This assumption will have to be tested, however, using paradigms which vary the difficulty of identifying stimulus saliency based on task set foci.

Finally, trial numbers for reversals were too low (40 reversals) to assess an effect of reversals on subsequent memory. However, no reliable difference in memory was observed when comparing memory on three trials before reversals with memory on three trials after reversals [younger adults: $t(1,27) = -0.33$, $P = 0.74$; older adults: $t(1,21) = 0.34$, $P = 0.73$].

Relationship of Pupil Responses to Losses and LC Integrity as Well as Memory. There was a trend for older adults with higher NM signal intensity to show larger pupil diameters to loss feedback (Fig. S3; older adults: $r = 0.44$, $P = 0.08$; younger adults: $r = 0.07$, $P = 0.75$). This suggests that older adults with higher LC integrity show stronger responses to salient stimuli. However, interindividual differences in pupil diameter to loss feedback did not explain a significant amount of variance in the increased memory related to loss feedback (older adults: $r = 0.21$, $P = 0.45$; younger adults: $r = -0.01$, $P = 0.97$). Similarly, the relationship between NM signal intensity in the LC and emotional memory in older adults (cf. Fig. 2*E*) was not reliably diminished when partialling out interindividual variance related to pupil diameters during loss feedback (older adults: $r = 0.53$, $P < 0.05$). Also, the trend of the relationship between memory data and pupil data did not survive correction for multiple comparisons across the three correlations. However, it should be noted that we could only use data from a reduced sample (25 younger adults and 16 older adults) to assess interindividual differences in pupil diameters, as pupil measurements in the scanner were more prone to noise during the acquisition. It is conceivable that the smaller power in the resulting measurements did not allow for reliable detection of small- or medium-sized effects related to interindividual differences in pupil diameter.

Similarly, although we did not observe a reliable interaction of condition and age group [$F(1,76) = 1.11$, $P = 0.33$], when inspected separately per age group, the difference in PD between change and gain trials as well as loss and gain trials was reliable in older adults [$t(1,15) = 3.67$, $P < 0.05$ and $t(1,15) = 3.48$, $P < 0.05$] but not in younger adults [$t(1,15) = 1.68$, $P = 0.11$ and $t(1,15) = 0.83$, $P = 0.41$]. This does not provide statistical proof of a difference, although it is possible that with a larger sample an age interaction would have been reliable. This might point to a stronger top-down focus on reversal and loss feedback in older adults which might have increased subjective saliency of these events in older adults.

1. Nassar MR, et al. (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat Neurosci* 15:1040–1046.
2. Oostenveld R, Fries P, Maris E, Schoffelen JM (December 23, 2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*, 10.1155/2011/156869.
3. Yushkevich PA, et al. (2006) User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 31: 1116–1128.
4. Zou KH, et al. (2004) Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports. *Acad Radiol* 11:178–189.
5. Sasaki M, et al. (2006) Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in Parkinson's disease. *Neuroreport* 17:1215–1218.
6. Sara SJ, Bouret S (2012) Orienting and reorienting: The locus coeruleus mediates cognition through arousal. *Neuron* 76:130–141.
7. Hämmerer D, et al. (2017) Emotional arousal and recognition memory are differentially reflected in pupil diameter responses during emotional memory for negative events in younger and older adults. *Neurobiol Aging* 58:129–139.
8. Hedden T, et al. (2012) Failure to modulate attentional control in advanced aging linked to white matter pathology. *Cereb Cortex* 22:1038–1051.

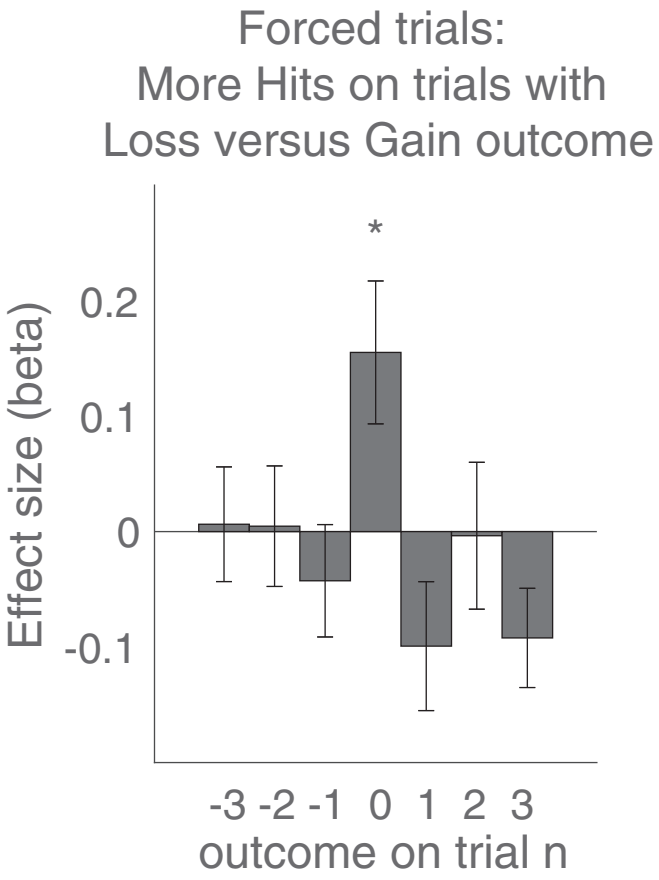

Fig. S1. Effect sizes of predicting memory success based on receiving a gain or a loss feedback on current, preceding, and following trials for forced trials only (for all trials, see main manuscript). Only losses on the current trial ($n = 1$) were predictive of memory success. *$P < 0.05$.

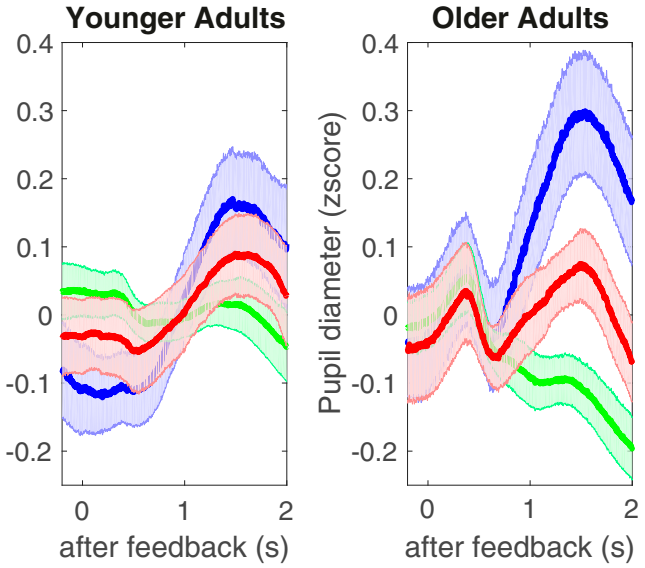

Fig. S2. Pupil diameters to loss and gain feedback and to reversal feedback (including gains and losses). Blue indicates reversals; red indicates loss; and green indicates gain.

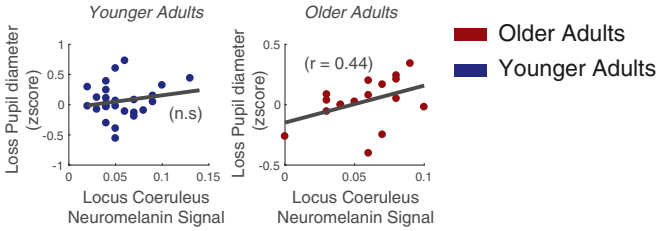

Fig. S3. Trend for pupil diameters to loss stimuli to be larger in those older adults with higher NM signal intensity in the LC.

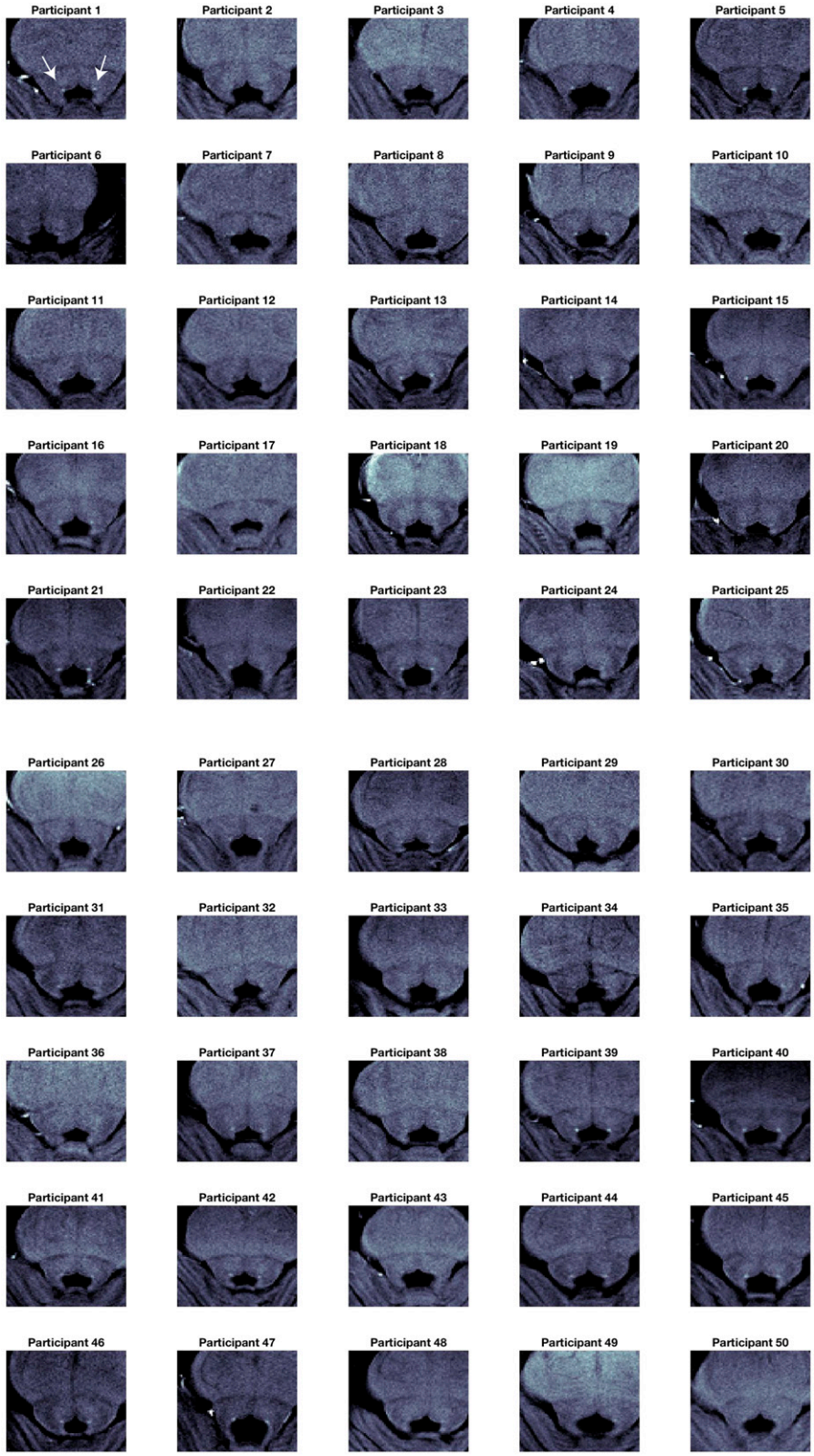

Fig. S4. Individual T1-weighted neuromelanin-sensitive scans as used for manually drawing individual masks (shown is the slice in the middle of the LC mask in axial view). Left and right LC are visible as hyperintense areas at the dorsal border of the pons just above the fourth ventricle (arrows in first participant). Participants 1 to 28 are younger adults.

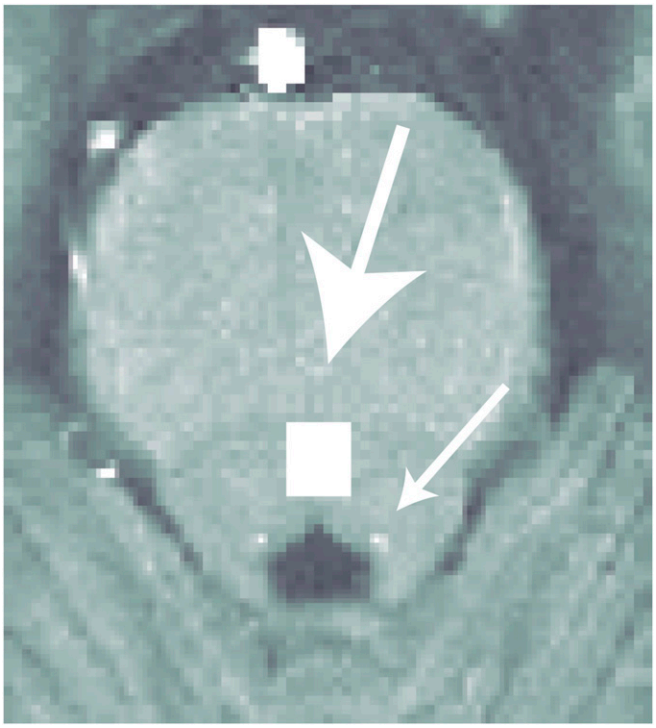

Fig. S5. Example participant showing the position of the mask for the control area in the pons (larger arrow on the white square; right LC is indicated with the small arrow).

Table S1. Sample description

Age group	Sample size, *n* (sex)	Mean age, y (SE)	Raven's matrices (SE)
Younger adults	28 (12 male)	23.14 (0.60)	16.26 (0.31)
Older adults	22 (10 male)	67.68 (1.21)	14.09 (0.52)

As a measure for fluid intelligence, a shortened version of Raven's matrices was used. Values indicate correct responses out of a total of 20 matrices.