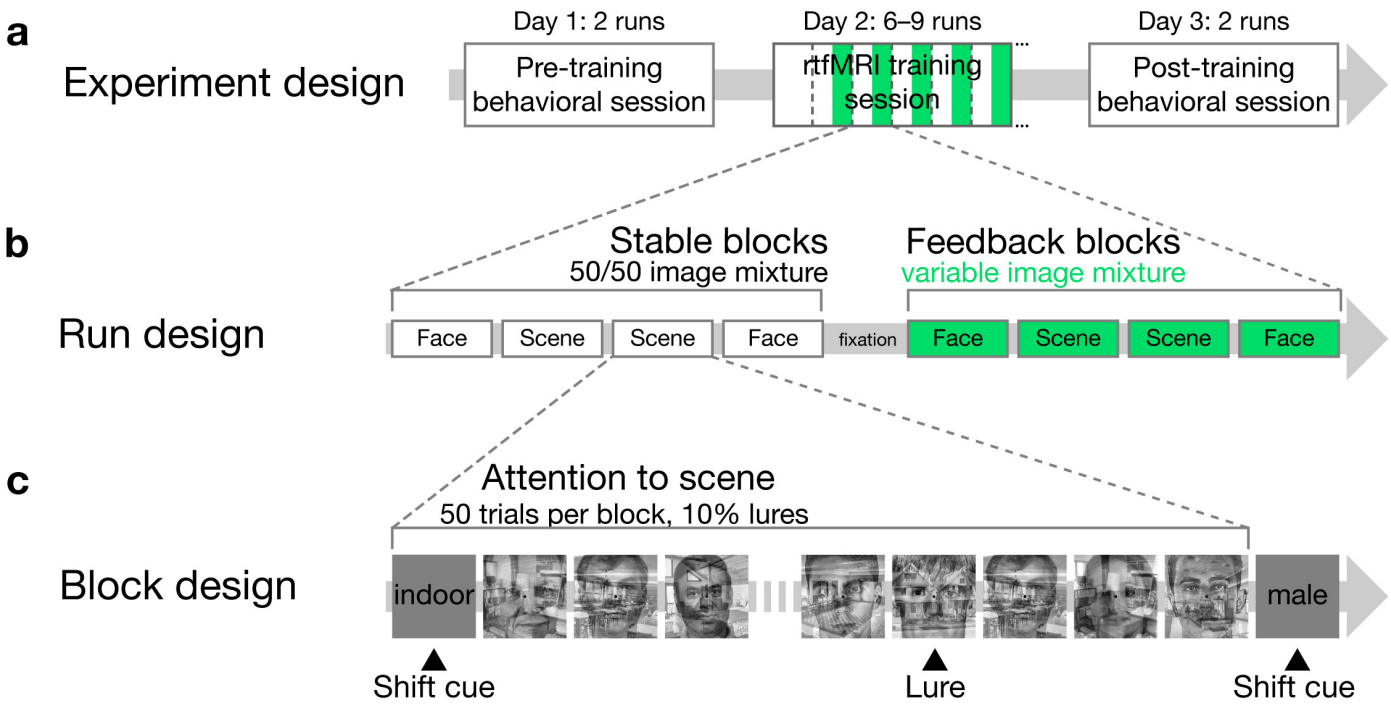
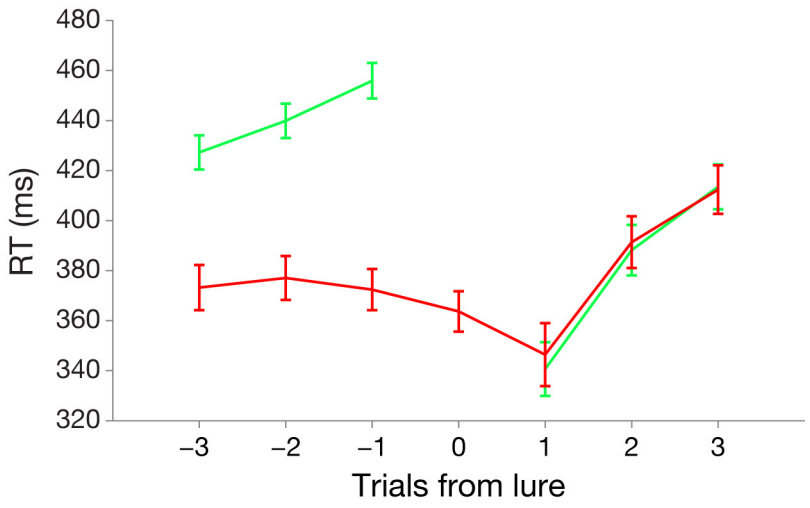


Supplementary Figure 1

Study procedure

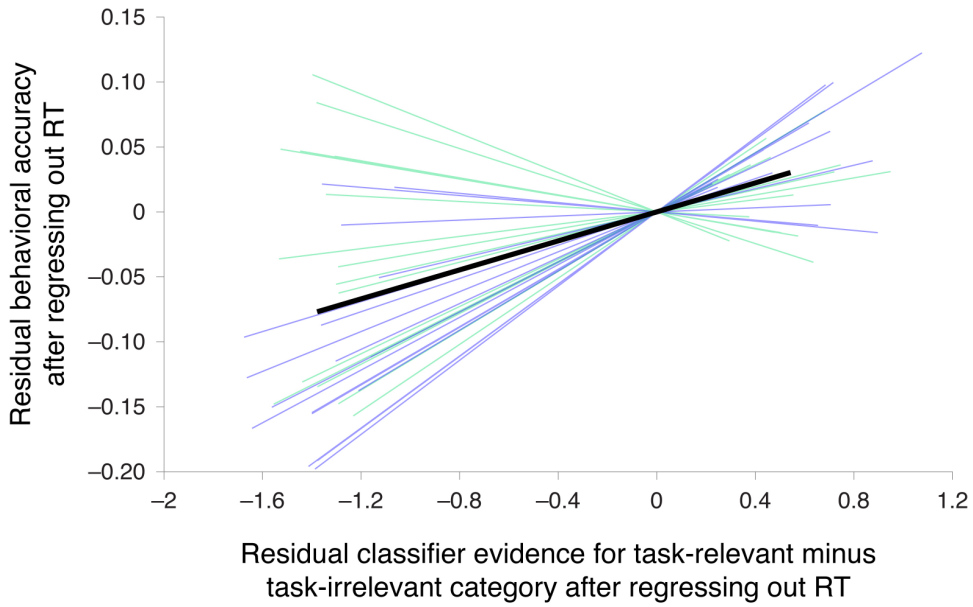
(a) Participants completed three sessions on different days, with different numbers of task runs. (b) Each run contained eight blocks of the sustained attention task. The pre-training, post-training, and stable blocks of the rtfMRI training sessions contained composite stimuli with an equal mixture of faces and scenes. In the feedback blocks of the rtfMRI session, the mixture of images was determined by real-time analysis of brain activity. (c) Each block began with a cue (1 s) that indicated the attended category (e.g., scene) and target subcategory (e.g., indoor). This was followed by a brief fixation period (1 s) and then 50 sequential stimuli (1 s each, no interstimulus interval), of which 90% were targets and 10% were lures. The blocks ended with a fixation period (4–6 s).



Supplementary Figure 2

Average RTs surrounding lures

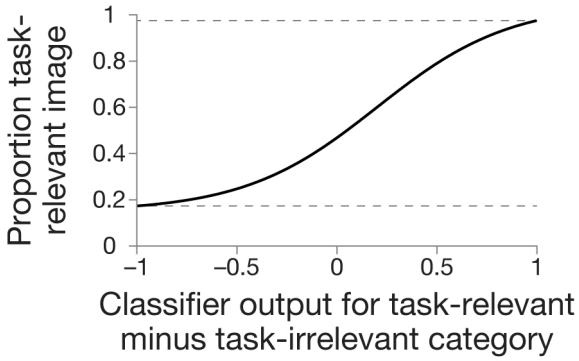
Green lines correspond to trials around correct rejections (CRs), where there was no behavioral response to the lure (presented at time = 0). Red lines correspond to trials around false alarms (FAs), where participants mistakenly responded to the lure. RTs were significantly slower prior to CRs than FAs (all timepoints, *p*s < 0.00001), consistent with the idea that FAs occurred when participants started responding habitually and were less attentive to the task. Error bars represent +/–1 s.e.m.



Supplementary Figure 3

Removing the influence of RT

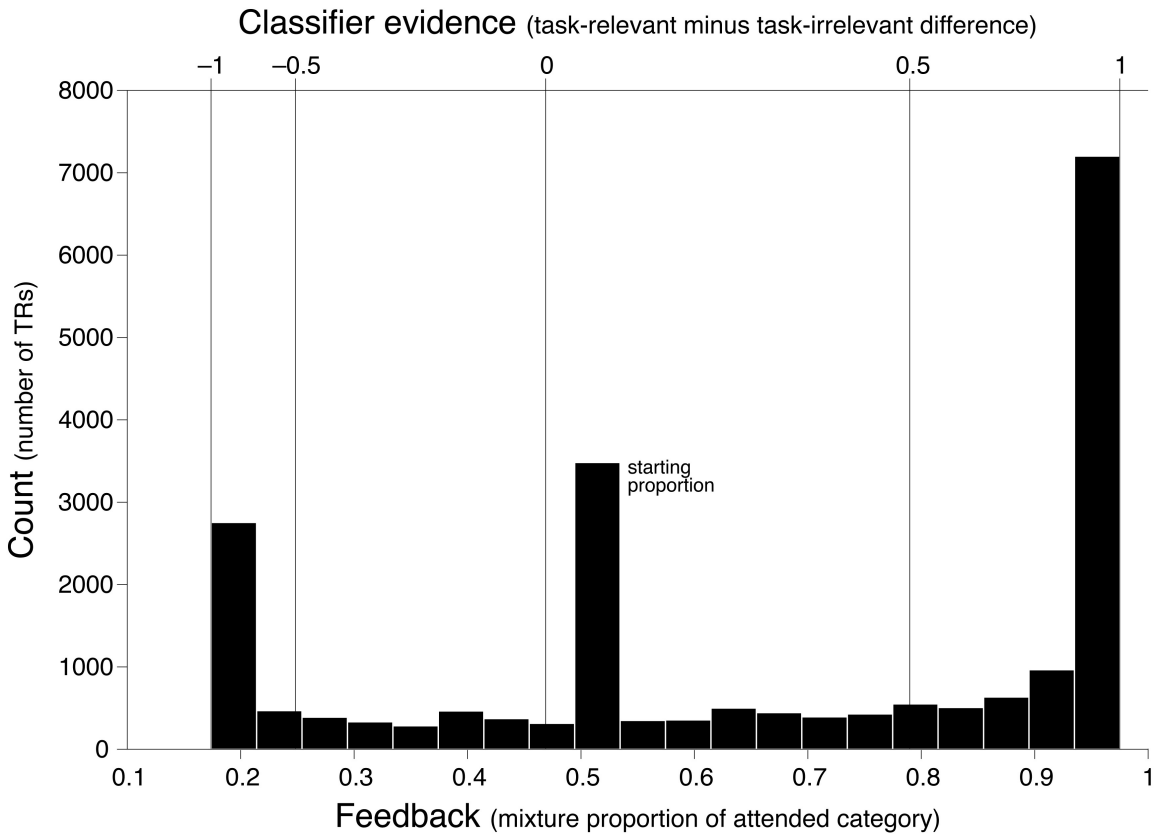
Both RT and classifier evidence for the task-relevant vs. task-irrelevant category were higher preceding CRs. To verify that the classifier was predictive of behavioral accuracy and not merely correlated with RT, we regressed RT out of classifier output and behavioral accuracy across trials and performed a partial correlation analysis. Green lines correspond to regression fits for feedback participants, and blue lines correspond to regression fits for control participants. The black line is the average fit. The relationship between classifier output and behavioral accuracy remained robust ($p < 0.00001$).



Supplementary Figure 4

Classifier-to-stimulus transfer function

The volume-by-volume classifier output for the task-relevant minus task-irrelevant category was mapped to the proportion of the image from the task-relevant category using a sigmoidal function. The inflection point on the classifier axis was set to 0.60, based on the average decoding accuracy in a pilot study. Given the nonlinearities in the function, this helped calibrate the feedback to a more sensitive range of classifier values. Image proportion ranged from 0.17 to 0.98, preventing the task-relevant image from ever disappearing completely, and providing a foothold for recovery from a serious lapse.



Supplementary Figure 5

Histogram of feedback

The bottom x-axis refers to proportion of the image from the task-relevant category in each composite stimulus. The y-axis refers to the number of TRs that contained stimuli with this mixture (in bins of width = 0.04) across all feedback blocks from all participants in the feedback group. The top x-axis depicts the correspondence between classifier output and feedback (computed using the transfer function in **Supplementary Fig. 4**). The most frequent values were in the highest and lowest bins, as well as in the bin including feedback of 0.5. The highest bin reflects cases in which the image from the task-relevant category was 98% of the composite stimulus. The lowest bin reflects cases in which the image from the task-irrelevant category was 83% of the composite stimulus. The bin with 0.5 was frequent because every block began with an equal mixture of the images from the task-relevant and task-irrelevant categories.