Supportive online material

Supplementary behavioural results:

In a separate behavioural study, 35 subjects (mean age 30, range 20-40, 14 men) performed the task out of the MRI scanner in two sessions on the same day. Task stimuli, parameters and trials type modulations were the same as the main experiment. They were given a short training period with each trial type separately then intermixed, as in the main study. In one of the two sessions (330 trials each) performance was rewarded (10 pence for each correctly identified third successive target in a modality), whereas in the other session it was not. Half the participants performed the rewarded session first, followed by the non-rewarded session, and in the other half of the participants this order was reversed. For this latter group, during the first non-rewarded session, they were not informed about the possibility of rewards in the subsequent session. The order manipulation was conducted to examine the relationship between incentive effects and practice. The reward consisted of a monetary bonus (10 pence) and the cash-register 'ka-ching' played following the correct detection of three successive targets.

As in the main experiment, there was evidence for graded cognitive set shift towards reward relevant dimensions. For both verbal and spatial domains, when the current trial was a target in that domain, performance was improved according to the number of previous targets in that domain, in terms of both faster reaction times successive targets lead to shorter reaction times (RT-spatial $F[3,31]=66$, $p<0.001$; verbal $F[3,31, 43$, $p<0.001$) and greater accuracy (Acc spatial $F[3,31]=33$, $p<0.001$; verbal $F[3,31]=33$, $p<0.001$). This effect also interacted with the type of trial in the non-target dimension (AY, BX, BY) (RT spatial $F[2,32]=43$, $p<0.001$; RT verbal $F[2,32]=91$, $p<0.001$) with longer RTs longer when the non-target dimension was of the AY type). Subjects were again not merely more generally attentive to the stimuli, but their attention became selective for one dimension with diminishing sensitivity to trial type differences in the alternate modality. This was revealed by the significant interaction in RT between the reward expectation from a given dimension (BY, AX^1, AX^2, AX^3) and trial type in the second dimension (BY, BX, AY), for both spatial and verbal dimensions (RT spatial $F[6,28]=5$,$p<0.001$ RT verbal $F[6,28]=4$,$p<0.005$).) Similar effects were found for accuracy, with a larger difference in accuracy between first and third targets when the non-target trial type was AY than BY (Acc spatial $F[6,28]=7$, $p<0.001$; RT verbal $F[6,28]=4.5$, $p<0.01$).

To explore the effects of monetary reward, a second ANOVA was undertaken using within subject factors: early vs late performance of the task incorporating practice effects; early vs late information about monetary reward; target trial type (BY, AX^1, AX^2, AX^3) and modality. On subjects

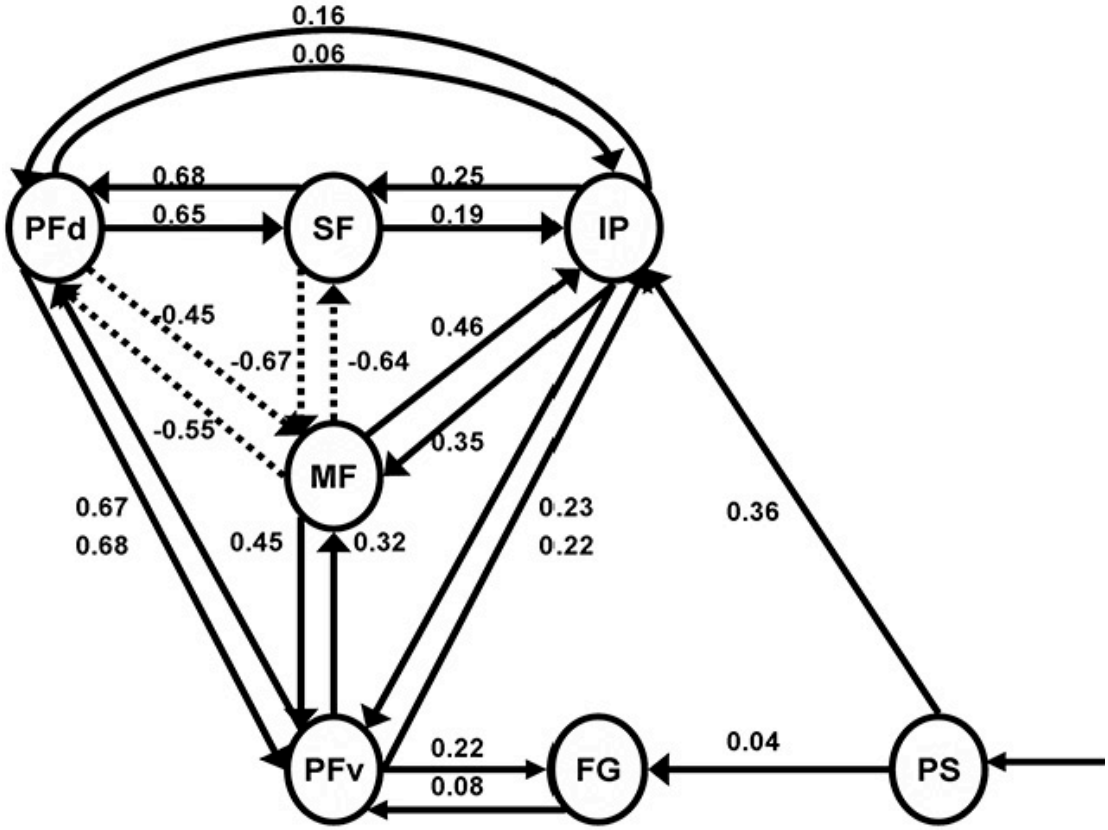
trained without initial monetary reward, the delayed introduction of reward lead to a greater increase in accuracy than further practice by subjects with early monetary reward (RT main effect of early vs late performance $F[1,32]=27, p<0.001$; no main effect of early reward $F[1,32]=0.1$, ns; main effect of target trial type $F[1,32]=25$, $p<0.001$; interaction between early vs late performance and early vs late reward $F[1,32]=5$, $p<0.05$; interaction between early vs late performance and trial type $F[1,32)=22$, $p<0.001$; no three-way interaction between early vs late performance, early vs late reward and trial type $F([1,32]=0.1$, ns).

Supplementary neuroimaging results

Trials in which reward was achieved (i.e., third successive target trials) resulted in an additional activation within medial frontal cortex, located more caudally to the reward expectation region (peak 6,30,46, $t=3.87$, FDR<0.01, cluster 255 voxels). We did not find reward expectation activation in orbital or ventrolateral prefrontal cortex.

During the trial sequences, the cognitive set biases were usually established by one or more dimension-specific target trials following one or more neutral (non-target) trials. However, we also included events in which a target in one dimension was followed by a target in the other dimension, requiring subjects to break an established cognitive set bias in favour of the previously less relevant set. This change does not reverse the reward relevance of letters or locations, but may involve a set shift between dimensions. For example, a first spatial target AX-$_{\text{spatial}}$ may be followed by a first verbal target AX-$_{\text{verbal}}$.. The spatial bias index is reset to zero and the verbal bias index set to one. This is important because only with this extradimensional shift is there a set-transition without a significant reduction in the conflict between the two sets (both have level one bias in this example). A discrete activation in the left anterior inferior frontal gyrus (BA 47) was observed in association with this extradimensional shift (-42, 44, 2, $t=4.16$, $p<0.001$ uncorrected, cluster 46 voxels) consistent with previous imaging data of extradimensional shifts (Hampshire and Owen 2006, Rogers, Andrews, Grasby, Brooks and Robbins 2000). Despite this role in set reversal, it its noteworthy that this region is not significantly activated in association with establishing cognitive set *per se* in response to reward expectation, or preferentially active for establishing one or other set.

Supplementary Figure S1. The preferred dynamic causal model included medial frontal (MF) cortex, dorsal (PFd) and ventral (PFv) lateral prefrontal cortex, superior frontal sulcus (SF), intraparietal cortex (IP), fusiform gyrus (FG) and prestriate cortex (PS). The intrinsic connections (from DCM matrix A) are given for positive (solid lines) or negative (dashed lines) connections with the connectivity coefficient values (Hz).



Supplementary Table S1. Foci of significant activation associated with increasing bias towards verbal or spatial dimensions, corresponding to a positive or negative correlation between BOLD signal and the parametric regressor of bias. Multiple peaks are shown for a cluster if >8mm apart. Activations are significant at FDR $p<0.05$ for whole brain comparisons for spatial bias, and FDR $p<0.05$ within a reduced search volume of area 45 and fusiform gyrus for verbal bias (see methods).

Region	cluster size	t-value	coordinate (MNI space) x	y	z
Spatial Bias					
Intraparietal cortex	1524	7.36	-14	-52	54
	"	6.77	-22	-60	54
	760	7.33	20	-62	60
Sup. Frontal sulcus	431	5.88	-20	-2	54
lateral occipital sulcus	332	5.63	42	-60	2
Lingual gyrus	302	5.54	18	-78	-8
PCS-SFS junction	216	4.9	24	-4	54
post central sulcus	30	4.54	28	-32	50
middle occipital gyrus	43	4.35	28	-84	16
	29	4.16	24	-80	42
	2	3.71	32	-78	30
Verbal Bias					
Inf. Frontal gyrus	128	4.77	-50	38	8
	"	4.71	-54	34	0
fusiform gyrus	2	3.75	-38	-62	-20