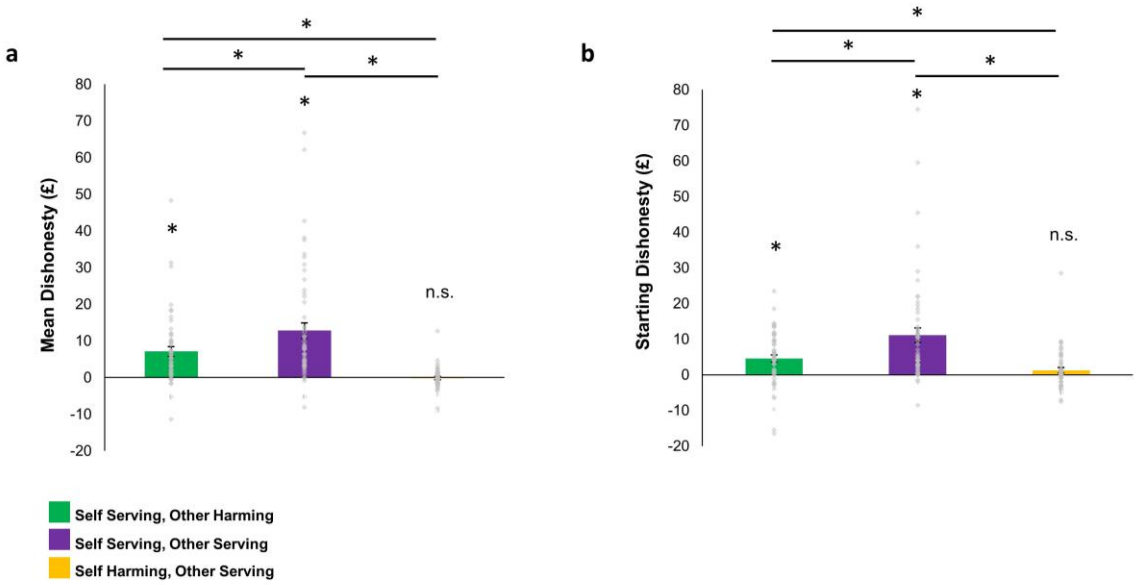
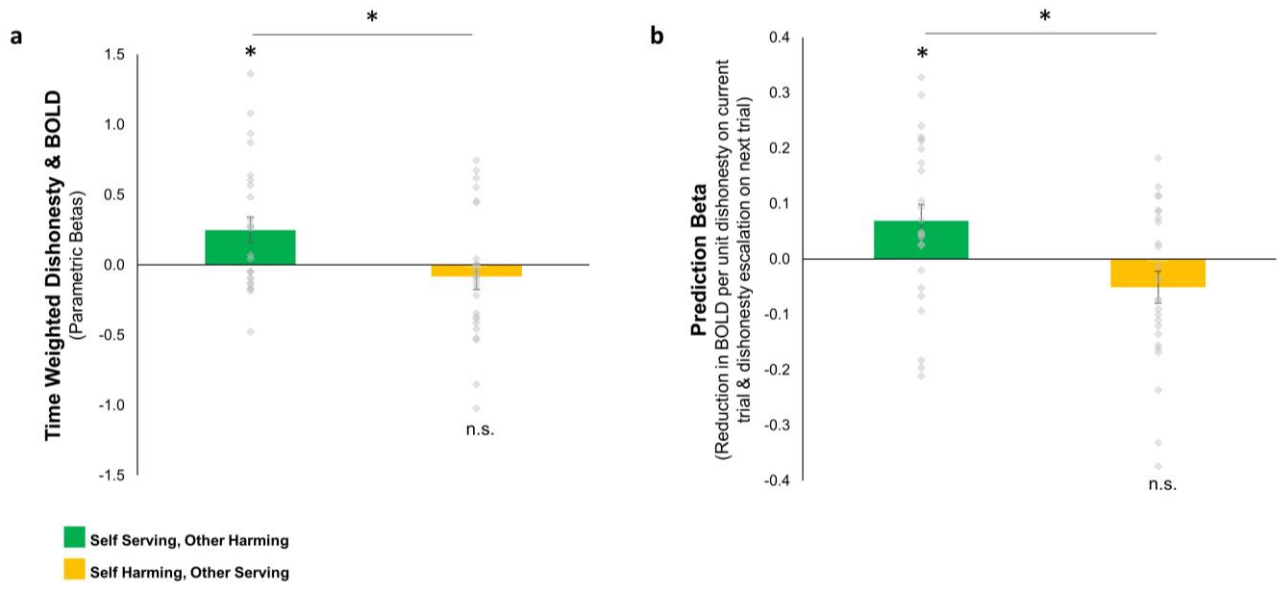


Supplementary Figure 1

Mean Dishonesty and Starting Dishonesty

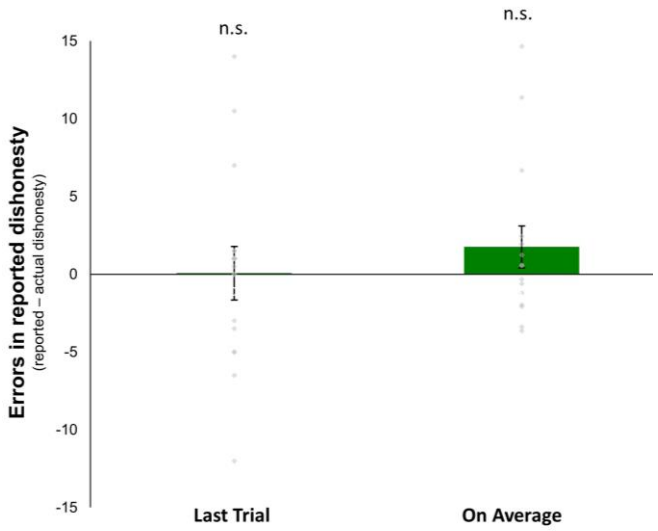
(a) Mean dishonesty over the course of the block (Self-Serving-Other-Harming: t_{54} = 5.49, p<0.001; Self-Serving-Other-Serving: t_{54} = 6.16, p<0.001; Self-Harming-Other-Serving: t_{54} = -0.39, p=0.70 - one sample ttests vs 0; comparisons between conditions: Self-Serving-Other-Harming vs Self-Harming-Other-Serving: $F_{1,53}$ = 26.44, p<0.001; Self-Serving-Other-Serving vs Self-Harming-Other-Serving: $F_{1,53}$ = 39.67 p<0.0001; Self-Serving-Other-Serving vs Self-Serving-Other-Harming: $F_{1,53}$ = 7.72, p=0.008 - statistics reported for separate 2 way ANOVAs with condition as a 2 level repeated factor controlling for study). **(b)** Starting dishonesty (Self-Serving-Other-Harming: t_{54} = 4.48, p<0.001; Self-Serving-Other-Serving: t_{54} = 5.56, p<0.001; Self-Harming-Other-Serving: t_{54} = 1.69, p=0.097, one sample ttests vs 0; Self-Serving-Other-Harming vs Self-Harming-Other-Serving condition: $F_{1,53}$ = 7.74, p=0.007; Self-Serving-Other-Serving vs Self-Harming-Other-Serving condition: $F_{1,53}$ = 21.99, p<0.001; $F_{1,53}$ = 8.84, p=0.004, 2 way repeated measure ANOVAs controlling for study). N=55.
Error bars represent standard error of the mean.
*p<0.05, n.s. = non-significant



Supplementary Figure 2

Voxels restricted to bilateral amygdala

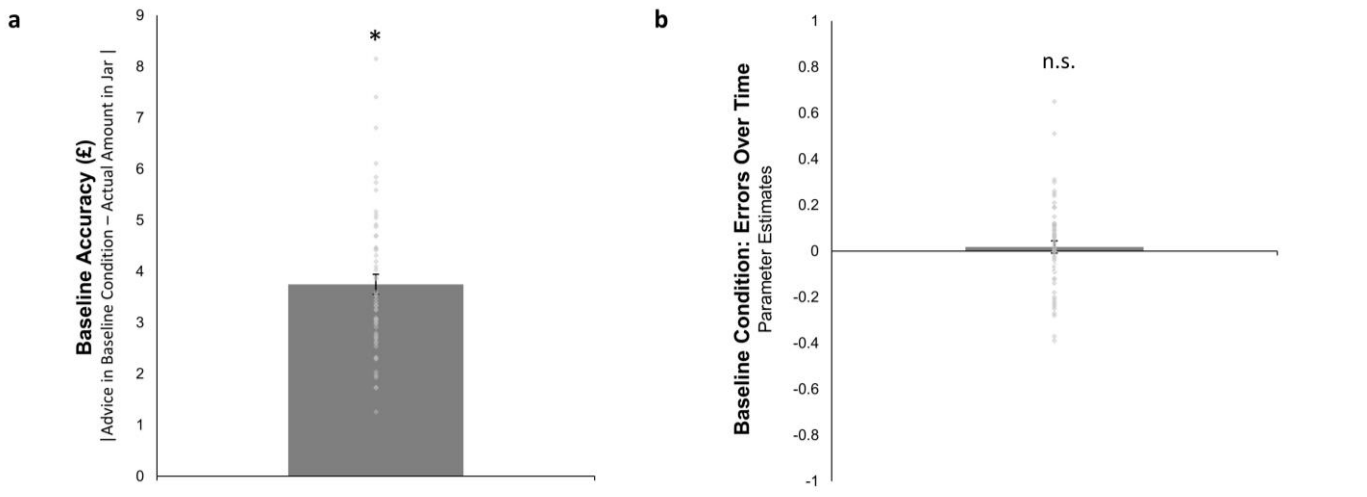
fMRI analysis was repeated on voxels in our ROI that were restricted to the anatomically defined amygdala. This analysis generated similar results as those portrayed in Figures 3-4. Specifically, **(a)** time-weighted dishonesty regressor positively correlated with BOLD response when dishonesty was Self-Serving-Other-Harming (t_{24} = 2.68, p=0.01, one sample ttest vs 0) but not when it was Self-Harming-Other-Serving (t_{24} = -0.9, p=0.38, one sample ttest vs 0) with the former parameter betas significantly greater than the latter (t_{24} = 2.60, p=0.02, paired sample ttest). **(b)** Reduction in BOLD response to one unit dishonesty on a current trial relative to the last predicted dishonesty escalation on next trial relative to current trial, when dishonesty was Self-Serving-Other-Harming (t_{24} = 2.30, p=0.03, one sample ttest vs 0) but not when Self-Harming-Other-Serving (trend in opposite direction t_{24} = -1.77, p=0.09, one sample ttest vs 0) with the former betas significantly larger than the latter (t_{24} = 3.05, p=0.01, paired sample ttest). N=25.
Error bars represent standard error of the mean.
*p<0.05, n.s. = non-significant



Supplementary Figure 3

Awareness of dishonesty

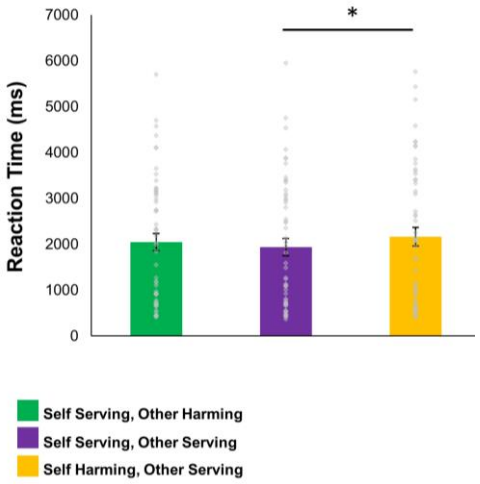
Fifteen participants, who completed the task outside of the scanner in testing cubicles in Experiment 1, were also asked immediately after the Self-Serving-Other-Harming block of trials to estimate the magnitude by which they gave advice over and above what they actually thought was in the jar on the last trial, as well as on average throughout that block (order of these two questions was counterbalanced). Participants did not know in advance that they would be asked to do this. Comparing participants' self-reports to their actual dishonesty revealed no significant differences on the last trial (t_{14} = 0.04, p>0.95, one sample ttest against a test value of 0) and on average (t_{14} = 1.29, p=0.22, one sample ttest against a test value of 0).
n.s. = non significant



Supplementary Figure 4

Baseline Accuracy

(a) Mean absolute error in the baseline condition (t_{54} = 19.06, p<0.001, n=55, one sample ttest vs 0). **(b)** Parameters show that errors were not changing with time in the baseline condition (regressing error rate over the 60 trials on trial number for each participant revealed no change over time: t_{54} = 0.67, p=0.51, n=55, one sample ttest vs 0). As participants did not receive any feedback it is not surprising that they did not show significant improvement over the course of the block.
Error bars represent standard error of the mean.
*p<0.05



Supplementary Figure 5

Reaction Time

RTs were slower in the Self-Harming-Other-Serving condition than the Self-Serving-Other-Serving condition (t_{54} = -2.19, p=0.03, n=55, paired sample ttest). Note, there was no correlation between mean RT and escalation of dishonesty in any of the conditions (all P > 0.3).
Error bars represent standard error of the mean.
*p<0.05