Supplementary Information

SI 1: Multivariate fMRI analysis

To identify brain regions in which we were able to decode the identity of the objects associated with the counterfactual action, we used multi-voxel pattern analysis (MVPA; Haxby et al., 2001; Mitchell et al., 2008, Norman, Polyn, Detre, & Haxby, 2006). To acquire a separate regressor for each critical trial, at the single subject level, the hemodynamic response function deconvolution for the 8-second period when participants were asked to make their judgments was once again performed using AFNI's 3dDeconvolve with a BLOCK model set for an eight-second event duration. This deconvolution was performed on the unsmoothed data. An ordinary least squares regression was performed with the onset of the 8-second decision period used as the primary regressor and motion parameters entered as regressors of no interest. The stim_times_IM argument was used to produce separate regressors for each trial.

Our objective was to classify which of the two possible objects served as the low-value alternative on each critical trial, and which of the two possible objects served as the high-value alternative on each critical trial. Prior to performing these classifications, we further divided the data based on whether a trial was an actual action or force trial, and whether we would be classifying an object associated with the rational or the irrational alternative action. Recall that low-value items serve as the rational alternative in the reject game and the irrational alternative in the keep game. Conversely high-value items served as the rational alternative in the keep game but as the irrational alternative in the reject game. Therefore, an item's value was orthogonal to its rationality. For example, for each subject's two low-value items (e.g. the ring and the stuffed animal) we divided the data into four conditions: a) actual action trials where the ring and stuffed animal were rational alternatives (when the subject was playing the reject game;

32 trials), b) actual action trials where the ring and stuffed animal were irrational alternatives (when the subject was playing the keep game; 32 trials), c) force trials where the ring and stuffed animal were rational alternatives (when the subject was playing the reject game; 32 trials), and d) force trials where the ring and stuffed animal were the irrational alternatives (when the subject was playing the keep game; 32 trials).

To identify brain regions in which we were reliably able to decode the object identity we conducted a whole-brain searchlight analysis (Kriegeskorte & Bandettini, 2007). Using the Searchmight Toolbox (Pereira & Botvinick, 2010), a cube with a 2-voxel (6mm) radius was centered at each voxel and Matlab's built-in Support Vector Machine (SVM) classifier was used to classify either the low-value (e.g. ring or stuffed animal) or high-value (e.g. the ax or the water filter) alternative objects on each trial. The classifier was trained to identify which of the two possible low or high-value objects was present on each trial (within one of the four groups of trials) using data from 15 runs, then tested on data from the 16th run. This leave-one-run-out procedure was iterated across runs so that each run was held out as the test run once. Once this procedure was completed, classification accuracies for each testing session were averaged together to produce a single whole-brain accuracy map for each of the four conditions (i.e., rational alternative in force trials, rational alternative in actual action trials, irrational alternative in force trials, and irrational alternative in actual action trials) for low-value items and a second set of accuracy maps for high-value items. This resulted in 8 accuracy maps for each subject. These accuracy maps were then averaged across value within each condition (for example subjects' low-value rational alternative in the actual action condition accuracy maps were averaged with subjects' high-value rational alternative in the actual action condition accuracy maps). This produced 4 accuracy maps per subject. Above chance accuracy maps were computed

by subtracting .5 from each voxel's accuracy score. These subject-wise above-chance accuracy maps were then smoothed by applying a 4 mm (2 voxel) Gaussian blur using AFNI's 3dMerge. The smoothed accuracy maps were then warped to TLRC space using ANTs. Next, we ran one-sample t-tests with AFNI's 3dttest++ to identify clusters of voxels where classification accuracy was significantly above chance across subjects. Finally, we used Monte Carlo simulation to perform cluster-wise corrections for multiple comparisons on the resulting t-maps by running the -Clustsim flag in 3dttest++.

SI 2: Multivariate fMRI results:

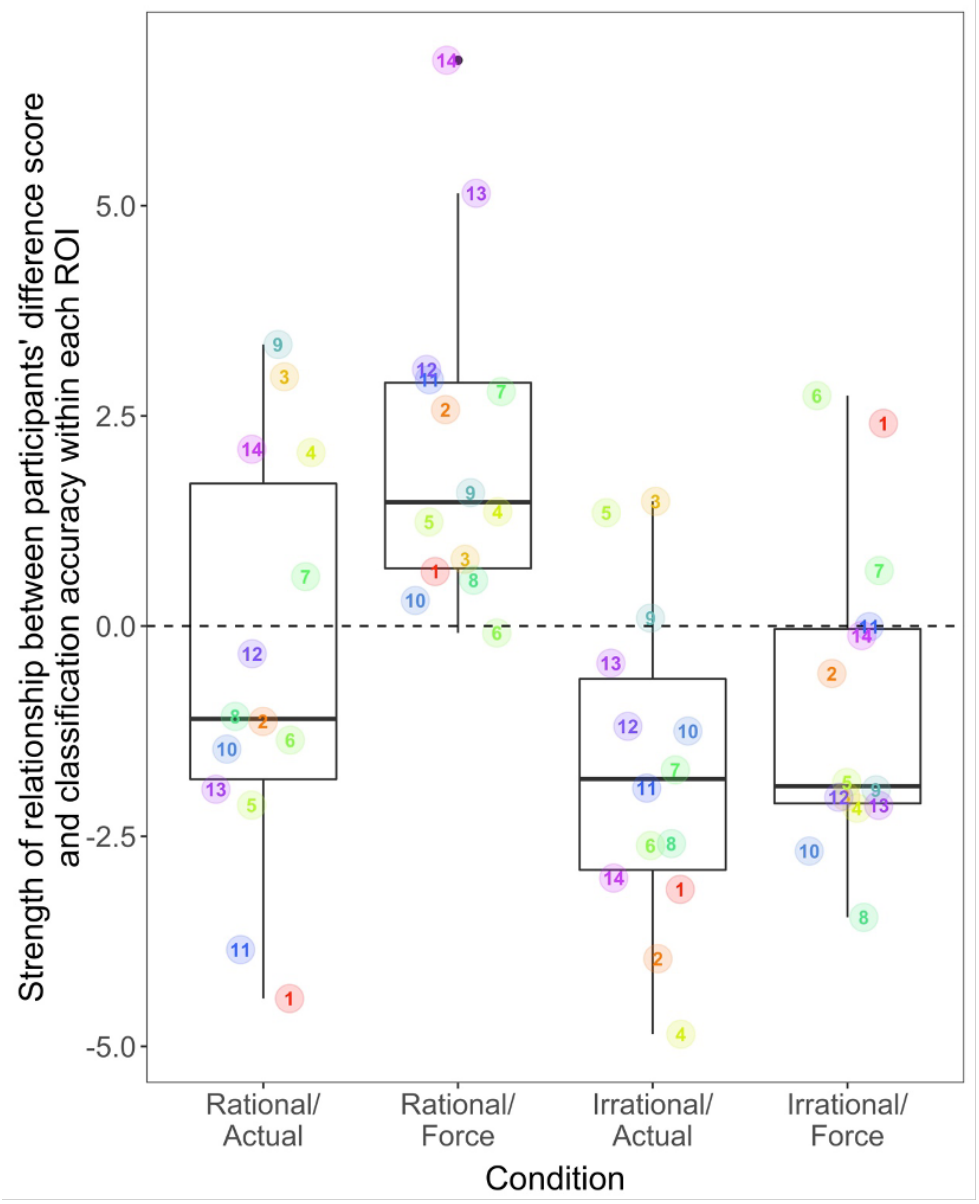
Our whole-brain searchlight analyses did not yield any clusters of voxels in which we were able to decode the identity of the rational or irrational alternatives significantly above chance accuracy. This was true for both the actual action and force conditions. Therefore, all of the subsequent analyses we conducted were within and across our 14 ROIs.

If the increased activation in our ROIs for force trials is related to consideration of counterfactual alternatives during force judgments, we would expect to be able to decode the identity of these alternatives. However, this should only be true when participants are making force judgments, and not when they are evaluating the agent's actual action. Critically, as predicted by prior research, we also expect that we should be able to decode the identity of the rational counterfactual action significantly better than the irrational counterfactual action during force judgments, since it is the rational counterfactual action that is relevant to determining one's force judgment.

Visual inspection of our results provides some support for this prediction (see SI Fig. 1). Specifically, we find that within each ROI, there tends to be a positive relationship between force judgments and classification accuracy, when classifying the rational alternative on force trials.

However, there tends not to be such a relationship for actual action trials (as opposed to force trials), or for the classification of irrelevant alternatives.

To test whether these trends are statistically significant we performed a linear mixed effects regression implemented with the *lmer* function in the *lme4* package in R (Bates et al., 2014) predicting classification accuracy from the interaction of condition (force vs. actual action), alternative (rational vs. irrational), and behavior (judgment). In this analysis, our categorical predictors were effect coded and we included random intercepts for subject and ROI. We then used the *anova* function in R to compare this model to a reduced model that excluded the three-way interaction. We found that the model that included the three-way interaction did not perform significantly better than the reduced model (X^2 (1, N = 36) = .413, p = .52). It is worth noting, though, that with only 36 participants we are likely underpowered to find a significant three-way interaction. In sum, significant caution is warranted in the interpretation of these results, but they may provide promising directions for further study.



SI Fig. 1. Average slopes from linear models predicting participants' difference scores from classification accuracy within each of our 14 ROIs. Slopes were averaged separately for each condition. Participants' difference scores best predict classification accuracy across ROIs when participants are making force judgments and a rational alternative is being classified.