

Supplementary Materials for

Neural pattern similarity unveils the integration of

social information and aversive learning

Irem Undeger^{1*}, Renée M. Visser², Andreas Olsson¹

¹ Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 9, 171 77 Stockholm, Sweden

² Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

*Corresponding author: irem.undeger@ki.se

Supplementary Material 1: Post-experimental questionnaire.

Answer the questions below, based on the first part of the experiment, when you were watching the other participants make choices.

Please make sure you answer every question. If you don't know, make a guess.



Did you receive any shocks when the above picture was chosen? If yes, how many?

How many times was the above picture chosen overall?

On a scale of 1-5, how much did you expect to receive shocks when the above picture was chosen?



Did you receive any shocks when the above picture was chosen? If yes, how many?

How many times was the above picture chosen overall?

On a scale of 1-5, how much did you expect to receive shocks when the above picture was chosen?



Did you receive any shocks when the above picture was chosen? If yes, how many?

How many times was the above picture chosen overall?

On a scale of 1-5, how much did you expect to receive shocks when the above picture was chosen?



Did you receive any shocks when the above picture was chosen? If yes, how many?

How many times was the above picture chosen overall?

On a scale of 1-5, how much did you expect to receive shocks when the above picture was chosen?

Supplementary Material 2: Post-experimental questionnaire.

Questionnaire - (one intentional / one unintentional player)

Please answer every question! Make a guess if you don't know. No right or wrong answers!

1. How did you experience the shocks from the player who *wanted* to shock you?

Somewhat
uncomfortable 1 2 3 4 5 Extremely
uncomfortable

2. How did you experience the shocks from the player who *did not want* to shock you?

Somewhat
uncomfortable 1 2 3 4 5 Extremely
uncomfortable

3. How many shocks did the player who *wanted* to shock you give you? Write a number.

4. How many shocks did the player who *did not want* to shock you give you? Write a number.

5. What do you think was the motivation of the person who decided to shock you?

6. How did you feel when you got shocks from the player who decided to shock you?

7. Would you like to give shocks the player who *wanted* to shock you? How many?

8. Would you like to give shocks to the player who *did not want to* shock you? How many?

9. Throughout the experiment, did you pay attention to who's making the choice? Did you think about the choice they made in the beginning (to give shocks or not)?

10. How angry did you feel when you received shocks from the player who *wanted* to shock you?

Not at all 0 1 2 3 4 5 Very

11. How angry did you feel when you received shocks from player who *did not want to* shock you?

Not at all 0 1 2 3 4 5 Very

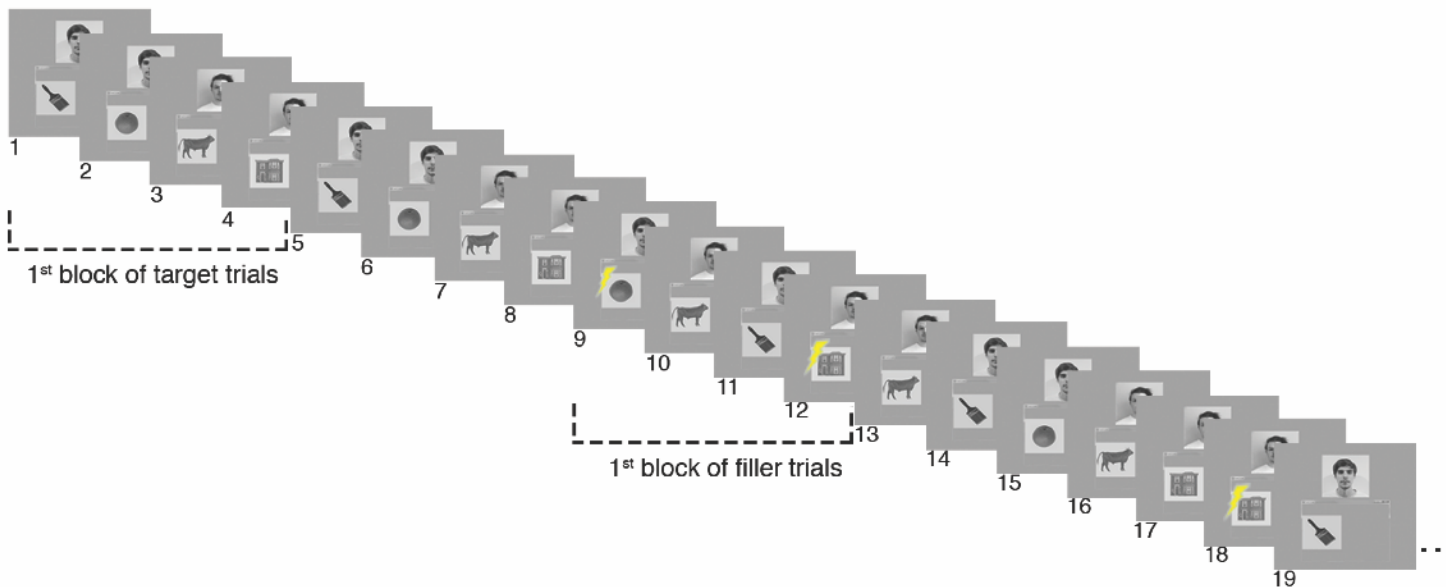
Supplementary Material 3: Instructions. A verbatim account of the instructions given to the participant about the experiment.

“Welcome to the experiment. I will now give a brief introduction about the experiment. Soon, you will be divided into two groups. We will use this lottery bag to decide which group you are in. There are two papers in this bag which say “Outside” and one that says “MR camera”. If you pick the one that says “MR camera”, you will be in the scanner for the rest of the experiment and I will be your researcher. This person will have an observer position. The other two people that are outside will be with my colleague and will be given a choice task. The people outside will have a choice to make in the beginning, that they cannot change later on. This is if they want to give shocks to the person that is laying in the MR scanner, or not. You are all free to answer yes or no, but depending on the answer the outcomes of your actions will change. Throughout the rest of the experiment, you will be making choices between two images. Each person will have their own two images and this will also not change throughout the experiment. In the case you would like to deliver shocks, the shocks will be delivered upon choosing one of the images and not the other. In the case that you would not like to deliver shocks, your connections to anyone else in this experiment will be lost.

What does this all mean for the person in the scanner? This person will have a passive role, meaning they will observe these decisions made by the other two people. This person needs to watch and learn if none or both of the other participants would like to give him or her shocks, and also which of the images are delivering shocks.

This will all be clearer once you have your assigned positions.”

Supplementary Material 4: Experimental design. The experiment consisted of repeated blocks of filler and target trials. The order of stimuli in target trials were fixed in order to ensure equal temporal distance in the trial-by-trial RSA, and these trials lacked electrical stimuli. Filler trials consisted of a semi-randomised order of stimuli, and included electrical stimulation in the case of an aversive choice. With the “target” and “filler” trial structure (below), we make sure that each correlation we report for a given trial pair in a condition, is separated by the same temporal distance. Namely, trial-to-trial correlations for each condition we compare are always separated by 3 other target trials, plus an average of 0-6 filler trials. In the trial scheme we provide below, you can see that the first trial of $CS_{+intent}$ and the second trial of $CS_{+intent}$ have 3 trials between them, and so does the first trial of $CS_{-intent}$ and the second trial of $CS_{-intent}$.

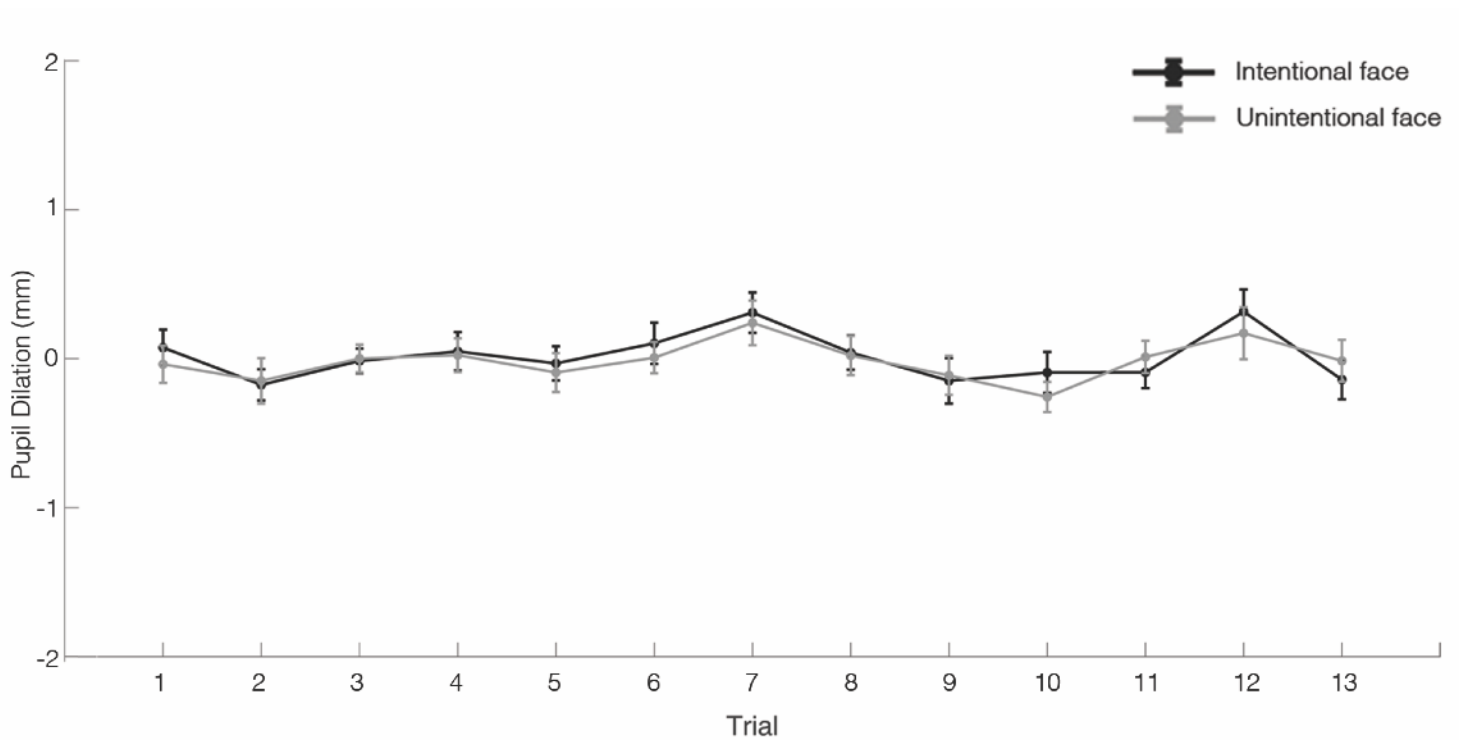


Supplementary Table 1: Covariate analysis of behavioural measures using reported believability of the interaction.

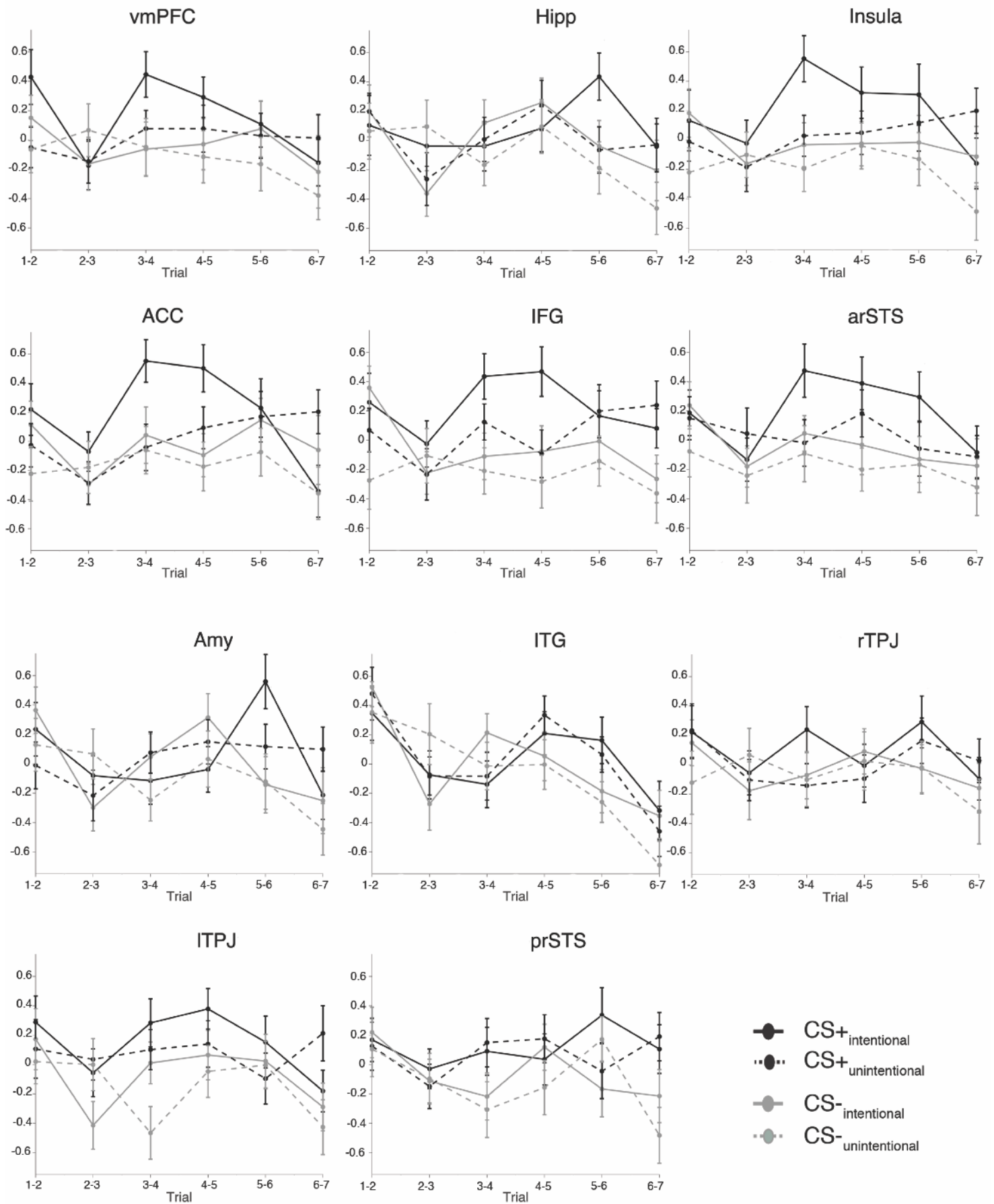
Rating	Intentionality (2)		Interaction (2x2)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
Discomfort	2.966	<i>0.096</i>	N.A.	N.A.
Number of shocks received	4.521	0.043	N.A.	N.A.
Revenge	3.036	<i>0.093</i>	N.A.	N.A.
Anger	13.368	0.001	N.A.	N.A.
Expectancy	5.121	0.032	12.474	0.039
Pupil dilation	1.313	0.261	1.752	0.195
Likeability	0.373	<i>0.547</i>	N.A.	N.A.

N = 28. All significant values ($P < 0.05$) are in bold, and those that are not significant in this analysis, but were before the believability adjustment, are in italics. NA = not available: some ratings were not collected for different types of CS's

Supplementary Figure 1: Pupil dilation responses to the faces of co-participants.



Supplementary Figure 2: Graphical representation of trial-by-trial neural pattern correlations in all of the ROI's used in the study.



Supplementary Table 2: Trial-by-trial RSA from fMRI data. Effects below the 0.05 threshold are highlighted in bold.

Region	Intentionality (2)		CS (2)		Interaction (2x2)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	5.797	0.0220	6.982	0.0126	0.004	0.945
Amygdala	0.852	0.363	1.902	0.177	0.277	0.602
Hippocampus	1.215	0.278	1.978	0.169	0.017	0.896
IFG	5.963	0.020	15.602	0.0004	0.001	0.970
ITG	0.085	0.771	0.808	0.375	0.237	0.629
Insula	5.422	0.026	7.893	0.008	0.009	0.921
arSTS	4.066	0.052	8.892	0.005	0.006	0.937
dmPFC	0.681	0.415	5.874	0.021	1.775	0.192
ITPJ	0.774	0.385	10.137	0.003	0.032	0.857
prSTS	0.493	0.487	5.443	0.026	0.022	0.882
rTPJ	0.708	0.406	2.346	0.135	0.070	0.792
vmPFC	3.610	0.066	4.458	0.042	0.244	0.624

All significant values ($P < 0.05$) are in italics, and those that reach FDR-corrected significance (for 12 ROIs, corrected per main or interaction effect) are in bold. ACC = Anterior Cingulate Cortex; IFG = Inferior Frontal Gyrus; ITG = Inferior Temporal Gyrus; arSTS = Anterior Superior Temporal Sulcus; dmPFC = Dorsomedial Prefrontal Cortex; ITPJ = Left Temporal Junction; prSTS = Posterior Superior Temporal Sulcus; rTPJ = Right Temporal Junction; vmPFC = Ventromedial Prefrontal Cortex.

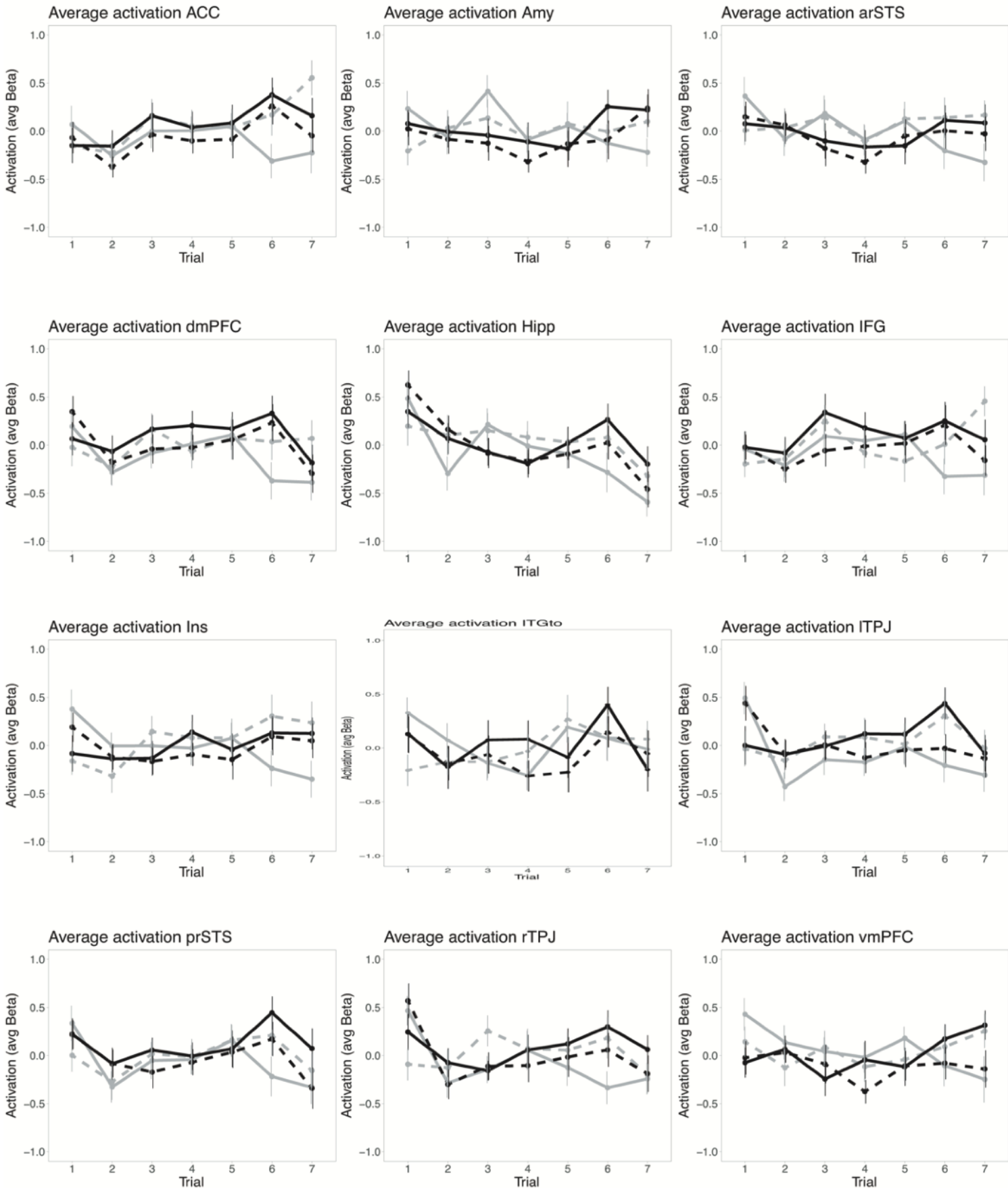
Supplementary Table 3: Covariate analysis of RSA correlations using reported believability of the interaction.

Region	Intentionality (2)		Interaction (2x2)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	4.912	0.035	0.270	0.607
Amygdala	4.311	<i>0.048</i>	0.150	0.702
Hippocampus	2.385	0.134	N.T.	N.T.
IFG	7.245	0.012	0.246	0.624
ITG	0.006	0.941	N.T.	N.T.
Insula	7.143	0.013	0.104	0.749
arSTS	1.263	0.271	N.T.	N.T.
dmPFC	1.680	0.206	N.T.	N.T.
ITPJ	1.123	0.299	N.T.	N.T.
prSTS	0.567	0.458	N.T.	N.T.
rTPJ	0.555	0.463	N.T.	N.T.
vmPFC	1.401	0.247	N.T.	N.T.

N = 28. All significant values ($P < 0.05$) are in bold. Effects that were significant in the main analysis but became non-significant via believability correction are in italics, effects that became significant via the correction are bold italic. NT = not tested: Areas without significant main effect of intentionality are not tested for an interaction with intentionality. ACC = Anterior Cingulate Cortex; IFG = Inferior Frontal Gyrus; ITG = Inferior Temporal Gyrus; arSTS = Anterior Superior Temporal Sulcus; dmPFC = Dorsomedial Prefrontal Cortex; ITPJ = Left Temporal Junction; prSTS = Posterior Superior Temporal Sulcus; rTPJ = Right Temporal Junction; vmPFC = Ventromedial Prefrontal Cortex.

Supplementary Figure 3: Average single-trial BOLD-fMRI activation in a priori ROI's.

- CS₊_{intentional}
- CS₊_{unintentional}
- CS₋_{intentional}
- CS₋_{unintentional}

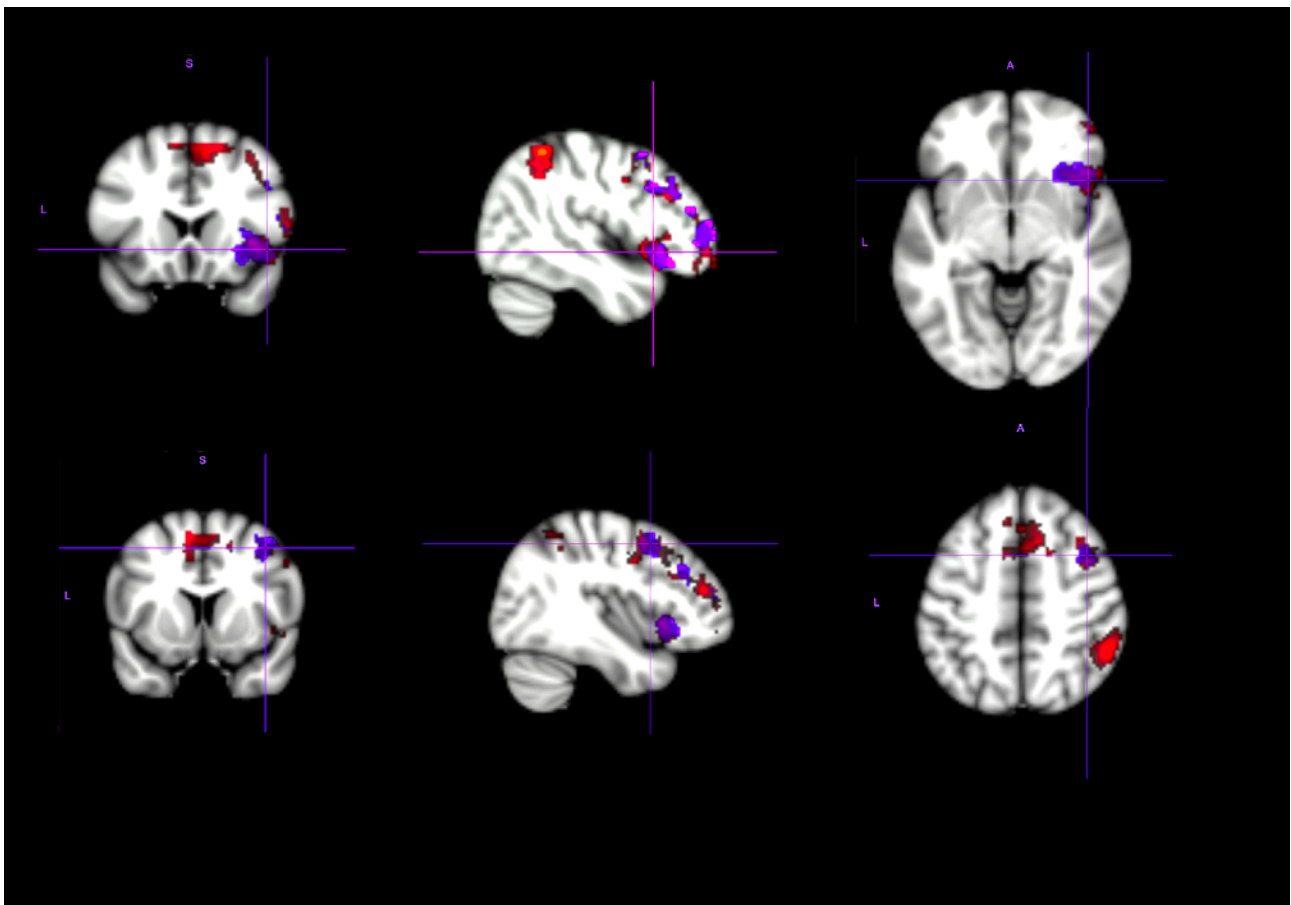


Supplementary Table 4: Average single-trial BOLD-fMRI activation in a priori ROI's.

Region	Intentionality (2)		CS (2)		Interaction (2x2)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	0.094	0.761	0.023	0.883	N.T.	N.T.
Amygdala	0.552	0.463	0.380	0.542	N.T.	N.T.
arSTS	0.072	0.790	1.483	0.232	N.T.	N.T.
dmPFC	0.074	0.787	3.088	0.088	N.T.	N.T.
Hippocampus	0.358	0.554	0.351	0.471	N.T.	N.T.
IFG	0.105	0.748	1.256	0.748	N.T.	N.T.
Insula	0.094	0.761	0.210	0.650	N.T.	N.T.
ITGto	0.953	0.336	0.193	0.663	N.T.	N.T.
ITPJ	0.302	0.586	1.363	0.252	N.T.	N.T.
prSTS	0.348	0.559	1.476	0.233	N.T.	N.T.
rTPJ	0.013	0.909	1.257	0.271	N.T.	N.T.
vmPFC	0.915	0.346	1.313	0.260	N.T.	N.T.

All significant values ($P < 0.05$) are in bold. NT = not tested: Areas without significant main effect of CS are not tested for an interaction with intentionality. ACC = Anterior Cingulate Cortex; IFG = Inferior Frontal Gyrus; ITG = Inferior Temporal Gyrus; arSTS= Anterior Superior Temporal Sulcus; dmPFC = Dorsomedial Prefrontal Cortex; ITPJ = Left Temporal Junction; prSTS = Posterior Superior Temporal Sulcus; rTPJ = Right Temporal Junction; vmPFC = Ventromedial Prefrontal Cortex.

Supplementary Figure 4: Overlapping activations for the manipulation of intentionality. Whole-brain univariate analyses for different conditions overlapped in one graph to facilitate the interpretation of results. In blue, the CS+>CS- contrast, in red the CS+_{intent}>CS-_{intent}. For the univariate analysis each GLM included target events, and 6 motion parameters. All lower-level analyses have been conducted with a voxel threshold of $z = 2.3$, $p = 0.05$. All higher-level analyses with a cluster threshold of $z = 2.3$, $p = 0.05$, using the mixed model, which uses FLAME 1 (FMRIB's local analysis of mixed effects) for the higher-level models. Thresholded statistical map overlaid on MNI 2mm template.



Supplementary Table 5: Univariate activation CS+>CS-. For methods, see

Supplementary Figure 4.

Region	X	Y	Z	Voxels	Z
R Frontal Pole	48	46	6	613	3,15
R Insular cortex	42	22	-6	519	4,19

All values are z-thresholded at 2.3. Anatomical labels are based on the Harvard-Oxford Structural Atlas. IFG = Inferior Frontal Gyrus; MFG = Middle Frontal Gyrus.

Supplementary Table 6: Univariate activation CS^{+intent} > CS^{-intent}. For methods, see Supplementary Figure 4.

Region	X	Y	Z	Voxels	Z
R MFG	40	6	50	1365	3,44
L SFG	-8	12	56	1186	3,55
R IFG	38	18	-2	793	3,8
R IPS	52	-48	48	725	3,66
-	-32	-66	-32	598	3,51

All values are z-thresholded at 2.3. Anatomical labels are based on the Harvard-Oxford Structural Atlas. MFG = Middle Frontal Gyrus; SMG = Supramarginal Gyrus; SFG = Superior Frontal Gyrus.

Supplementary Table 7: Summary of statistics of the fMRI data for the learning-phase ($n = 33$, 2x2 within-subjects ANOVA) in all ROIs derived from the Harvard-Oxford cortical and subcortical atlas.

Region	Intentionality (2)		CS (2)		Interaction (2x2)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
L ACC	2.155	0.152	1.966	0.170	NT	NT
R ACC	4.061	0.052	2.442	0.128	0.017	0.898
L Accumbens	0.473	0.497	0.505	0.482	NT	NT
R Accumbens	25.880	<0.0005	3.919	0.056	0.326	0.572
L Amygdala	0.696	0.410	5.120	0.031	0.598	0.445
R Amygdala	0.158	0.694	1.334	0.257	NT	NT
L Angular gyrus	1.019	0.320	5.064	0.031	0.485	0.491
R Angular gyrus	1.741	0.196	26.316	<0.0005	0.017	0.898
Brain stem	1.047	0.314	0.269	0.608	NT	NT
L Caudate	7.227	0.011	7.771	0.009	0.016	0.899
R Caudate	3.718	0.063	2.326	0.137	NT	NT
L Central operculum	2.547	0.120	4.570	0.040	0.737	0.397
R Central operculum	1.662	0.207	1.087	0.305	NT	NT
L Cuneus	0.113	0.738	4.042	0.091	NT	NT
R Cuneus	0.117	0.734	4.832	0.035	0.007	0.935
L Front operculum	8.643	0.006	9.017	0.005	0.041	0.842
R Front operculum	7.944	0.008	13.409	0.001	1.450	0.237
L Frontal pole	3.573	0.068	9.404	0.004	0.260	0.614
R Frontal pole	4.980	0.033	13.933	0.001	0.495	0.487
L Heschyl Gyrus	0.009	0.924	7.262	0.011	0.042	0.839
R Heschyl Gyrus	0.516	0.478	2.862	0.100	NT	NT
L Hippocampus	0.410	0.526	3.700	0.063	NT	NT
R Hippocampus	0.248	0.598	0.354	0.556	NT	NT
L IFG pars opercularis	11.134	0.002	0.643	0.429	0.214	0.647

Supplementary Materials

R IFG pars opercularis	3.659	0.065	4.559	0.041	0.203	0.656
L IFG pars triangularis	5.486	0.026	0.222	0.641	0.177	0.677
R IFG pars triangularis	1.930	0.174	10.983	0.002	0.100	0.754
L ITG anterior division	0.060	0.808	1.164	0.289	NT	NT
R ITG anterior division	0.600	0.444	1.864	0.182	NT	NT
L ITG posterior division	0.074	0.787	0.241	0.627	NT	NT
R ITG posterior division	0.937	0.340	0.610	0.441	NT	NT
L ITG temporooccipital	0.005	0.945	1.025	0.319	NT	NT
R ITG temporooccipital	0.054	0.818	0.066	0.799	NT	NT
L Insula	0.745	0.395	10.377	0.003	0.107	0.745
R Insula	2.572	0.119	11.228	0.002	1.180	0.286
L Intra calcarine sulcus	2.515	0.123	7.062	0.012	0.051	0.822
R Intra calcarine sulcus	1.544	0.223	3.785	0.061	NT	NT
L Juxtapositional lobule	5.923	0.021	4.986	0.033	0.842	0.366
R Juxtapositional lobule	5.080	0.031	6.265	0.018	0.144	0.707
L LOC inferior division	0.229	0.635	4.300	0.046	0.280	0.600
R LOC inferior division	1.952	0.172	<0.0005	0.988	NT	NT
L LOC superior division	0.109	0.743	6.021	0.020	0.027	0.870
R LOC inferior division	2.972	0.094	6.120	0.019	0.042	0.840
L Lingual Gyrus	13.146	0.001	2.130	0.154	1.429	0.241
R Lingual Gyrus	4.623	0.039	0.265	0.611	1.682	0.204
L Middle Frontal Gyrus	0.692	0.412	11.905	0.003	0.481	0.493

Supplementary Materials

R Middle Frontal Gyrus	0.819	0.372	17 846	<0.0005	0.018	0.893
L MTG anterior division	7.554	0.010	5.374	0.027	2.122	0.155
R MTG anterior division	1.374	0.250	0.364	0.551	2.985	0.094
L MTG posterior division	0.135	0.716	7.848	0.009	0.387	0.539
R MTG posterior division	2.418	0.130	11.073	0.002	0.768	0.387
L MTG temporooccipital	2.627	0.115	2.104	0.157	NT	NT
R MTG temporooccipital	0.943	0.339	6.007	0.020	0.273	0.605
L Orbitofrontal cortex	3.031	0.091	11.009	0.002	0.945	0.338
R Orbitofrontal cortex	4.080	0.052	11.893	0.002	0.750	0.393
L Occipital fusiform gyrus	1.375	0.250	1.583	0.217	NT	NT
R Occipital fusiform gyrus	0.626	0.435	0.105	0.748	NT	NT
L Occipital pole	0.552	0.463	3.163	0.085	NT	NT
R Occipital pole	0.590	0.448	0.437	0.513	0.381	0.541
L PCC	4.097	0.051	6.556	0.015	0.014	0.908
R PCC	8.863	0.006	11.356	0.002	0.007	0.934
L Pallidum	2.279	0.141	0.017	0.897	NT	NT
R Pallidum	0.766	0.388	0.079	0.781	NT	NT
L Paracingulate	1.327	0.273	8.711	0.006	0.942	0.339
R Paracingulate	3.020	0.092	16.688	<0.0005	0.549	0.464
L Parahippocampal ant.	0.554	0.462	0.823	0.371	NT	NT
R Parahippocampal ant.	0.944	0.338	0.131	0.720	NT	NT
L Parahippocampal post.	0.082	0.777	0.022	0.882	NT	NT

Supplementary Materials

R Parahippocampal post.	3.752	0.062	0.008	0.927	NT	NT
L Parietal operculum	1.200	0.282	0.766	0.388	NT	NT
R Parietal operculum	3.148	0.086	3.049	0.090	NT	NT
L Planum polare	0.454	0.505	1.079	0.307	NT	NT
R Planum polare	0.036	0.850	0.426	0.519	NT	NT
L Planum temporale	1.806	0.188	5.070	0.031	NT	NT
R Planum temporale	3.316	0.078	1.507	0.229	NT	NT
L Postcentral gyrus	4.086	0.052	8.085	0.008	0.189	0.667
R Postcentral gyrus	4.845	0.035	5.081	0.031	1.263	0.269
L Precentral gyrus	4.978	0.033	10.277	0.003	0.100	0.753
R Precentral gyrus	3.705	0.063	11.077	0.002	0.494	0.487
L Precuneous	1.627	0.211	5.781	0.022	0.052	0.821
R Precuneous	1.669	0.206	8.855	0.006	0.017	0.896
L Putamen	0.091	0.765	0.003	0.955	NT	NT
R Putamen	0.259	0.614	0.051	0.823	NT	NT
L Superior frontal gyrus	0.259	0.614	0.051	0.823	NT	NT
R Superior frontal gyrus	0.924	0.344	12.796	0.001	0.005	0.942
L SMG anterior division	0.187	0.668	1.452	0.237	NT	NT
R SMG anterior division	1.683	0.204	5.135	0.031	2.115	0.156
L SMG posterior division	1.734	0.197	4.510	0.042	0.283	0.599
R SMG posterior division	1.008	0.323	24.246	<0.0005	0.148	0.703
L Superior parietal lobule	0.124	0.728	5.081	0.031	0.234	0.632
R Superior parietal lobule	0.367	0.549	5.892	0.021	2.456	0.127
L STG anterior division	5.565	0.025	0.034	0.855	0.104	0.750

Supplementary Materials

R STG anterior division	0.366	0.549	4.086	<i>0.052</i>	0.131	0.720
L STG posterior division	0.832	0.369	13.367	0.001	0.611	0.440
R STG posterior division	0.447	0.509	11.794	0.002	0.360	0.553
L Subcallosal cortex	2.810	0.103	2.858	0.101	NT	NT
R Subcallosal cortex	3.126	0.087	1..858	0.182	NT	NT
L Supracalcarine sulcus	0.121	0.730	2.059	0.161	NT	NT
R Supracalcarine sulcus	0.171	0.682	2.589	0.117	NT	NT
L TFG anterior division	0.247	0.623	0.663	0.422	NT	NT
R TFG anterior division	0.196	0.661	0.220	0.642	NT	NT
L TFG posterior division	1.264	0.269	0.997	0.326	NT	NT
R TFG posterior division	0.032	0.860	0.060	0.808	NT	NT
L TFG occipital part	0.287	0.596	0.390	0.537	NT	NT
R TFG occipital part	0.451	0.507	0.560	0.460	NT	NT
L Temporal Pole	2.260	0.143	4.408	0.44	0.048	0.827
R Temporal Pole	0.802	0.377	6.077	0.019	1.095	0.303
L Thalamus	1.789	0.190	6.089	0.019	0.402	0.531
R Thalamus	0.022	0.884	0.570	0.456	NT	NT
L vmPFC	4.298	<i>0.046</i>	3.847	0.059	0.509	0.481
R vmPFC	3.961	0.055	4.503	0.042	0.293	0.592

All significant values ($P < 0.05$) are in italics, and those that reach FDR-corrected significance (for 111 ROIs, corrected per main or interaction effect) are in bold. NT = not tested: Areas without significant main effect of CS are not tested for an interaction with intentionality. ACC = Anterior Cingulate Cortex; IFG = Inferior Frontal Gyrus; ITG = Inferior Temporal Gyrus; LOC = Lateral Occipital Cortex; MTG = Middle Temporal Gyrus; ITG = Inferior Temporal Gyrus; PCC = Posterior Cingulate Cortex; SMG = Supramarginal Gyrus; STG = Superior Temporal Gyrus; TFG = Temporal Fusiform Gyrus, vmPFC = Ventromedial prefrontal cortex.

Supplementary Figure 5: A learning index was calculated by calculating per stimulus type the mean neural pattern similarity over trials 3-6 (after the first US-CS pairing) and subtracting the mean of the first two trials (before the first US-CS pairing, i.e., before the participant was able to connect the social information the CS's and their harm value; See methods for an overview of the design). Next, the difference between the CS_{+intent} learning index relative to the other indices was calculated, yielding one value per individual indicating the degree their brain showed evidence for integration of intentionality and aversive learning. Correlation of neural pattern correlations specific to intentional CS₊ and experienced shock strength from intentional vs. unintentional shocks. Conditioned responses are calculated by subtracting responses to the intentional CS₋, unintentional CS₊ and CS₋ from the intentional CS₊, after correcting for the habituation baseline. In the insula, the strength of pattern correlations in response to the intentional CS₊ (compared to all other stimuli) predicted the amount of discomfort experienced from the intentional shocks ($r = 0.69$, $p < 0.0005$), but not the unintentional ones ($r = -0.25$, $p > 0.1$), and predicted the difference between intentional and unintentional shocks ($r = 0.46$, $p = 0.009$). These findings should be interpreted with caution as the limited power of our sample size ($n=33$) did not allow for any firm conclusion about these relationships.

