Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/42/eaat4390/DC1

Supplementary Materials for

Does inappropriate behavior hurt or stink? The interplay between neural representations of somatic experiences and moral decisions

G. Sharvit, E. Lin, P. Vuilleumier, C. Corradi-Dell'Acqua*

*Corresponding author. Email: corrado.corradi@unige.ch

Published 16 October 2020, *Sci. Adv.* **6**, eaat4390 (2020)
DOI: 10.1126/sciadv.aat4390

This PDF file includes:

Supplementary Materials

Methods: Modelling whole-brain Signatures

We used Multivariate Pattern Regression to estimate a whole-brain model predictive of participants' unpleasantness in thermal and olfactory events. Consistent with previous implementations of this procedure (*24*, *25*), this analysis comprised the following independent steps.

Feature Selection. We identified inclusive masks, including only coordinates of theoretical interest. Consistent with Wager et al. (*24*), we selected coordinates that were preferentially associated with the term "pain" in the automated meta-analysis toolbox Neurosynth (*38*). In particular, we selected coordinates that were reported more often in articles about pain, as opposed to any other article (association test). This led to a mask derived by 512 published studies, comprising 18759 coordinates at a $2 \times 2 \times 2$ mm^3 resolution. Similar methods were used for identifying a mask of interest for olfactory unpleasantness. More specifically, we combined the association maps from the term "disgust" (103 studies), "olfactory" (74 studies) and "taste" (80 studies), leading to a mask of 4443 coordinates.

Data Selection. For our modelling purposes, we selected the data from an independent study (*18*). Although not focused on the mechanisms underlying moral cognition, this study shares the same set-up for thermal/olfactory stimulations of the present research, including comparable unpleasantness between pain and disgust, the presence of an anticipatory cue, and of a 3 sec countdown to constrain respiratory activity. Differently from the present research, Sharvit et al. (*18*) utilised three levels of unpleasantness, with LP and HP conditions associated with an intermediate temperature MP, and LD and HD associated with an intermediate odorant MD. Furthermore, in this previous study, an expectancy manipulation was implemented, with cues sometimes predictive of a different modality/unpleasantness with respect to the subsequent stimulus. For the purpose of the modelling, we reanalysed the data from the 20 participants from Sharvit et al.(*18*) through an *ad hoc* first level GLM characterised by two functional runs and 6 stimulations of interest (LP, MP, HP, LD, MD, and HD). Differently from the original analysis, here we did not consider congruency/incongruency with the preceding cue, but modelled conditions solely based on the bottom-up properties of the stimulations. In line with the present research, 6 runs out of 40 (20 subjects * 2 runs) were excluded to prevent discrepancies in unpleasantness between the two modalities. This led to 204 parameters: 102 for pain and 102 for disgust.

Data Reduction and Multivariate Modelling. We extracted the 102 parameters associated with pain data from the Neurosynth mask derived by pain studies, and the 102 parameters associated with disgust with the mask derived by disgust/olfaction/taste studies. This led to two data matrixes (102 parameters × 18759 coordinates and 102 × 4443 coordinates), each of which underwent dimensionality reduction using principal component analysis (PCA), to condense the large number of coordinates in a limited number of components that retained ~ 99.9% of variance of the original dataset. This allowed us to reduce the 18759 coordinates of the pain matrix into 95 components, and the 4443 coordinates in the disgust matrix in 96 components.

The scores estimated in the PCA were then fed into algorithms for multivariate regression to identify a model predictive of participants' unpleasantness ratings. In particular, we

employed an explorative approach, involving three well-known algorithms: Support Vector Regression (SVR) under radial basis kernel function, Random Forest regression (RF) and Least Absolute Shrinkage and Selection Operator (LASSO). For each of these algorithms we employed a leave-one-subject-out cross-validation approach, to assess the proficiency of the algorithm to predict heat/olfaction unpleasantness in an independent portion of subjects: e.g., a model trained on pain PCA scores in all-but-one subjects is used to predict pain unpleasantness in the remaining participant. For SVR and LASSO an additional nested cross-validation loop was included to identify at each fold the most suitable combination of hyper-parameters for the modelling (C, ε & γ for SVR and λ for LASSO). As an overall measure of predictive proficiency, we calculated the mean squared error (MSE), reflecting the deviation between unpleasantness actually rated, and the one estimated from brain activity. The resulting MSEs (one for heat and the other for olfaction) were considered to be significant if lower than the 5th percentile of the distribution of 1000 MSEs obtained by re-running the same analysis procedure on permuted datasets. SVR analysis was carried out using the LIBSVM 3.18 software (*39*). RF regression was carried out as implemented in the Matlab-based RF toolbox (https://code.google.com/archive/p/randomforest-matlab/), whereas LASSO was carried out using the *lasso* function from Matlab R2015b.

Overall, the unpleasantness associated with heat could be reliably predicted by all three algorithms, with SVR providing the smallest error (SVR: MSE = 173.97 [permutation-based cut-off = 244.19]; LASSO: 204.40 [233.57]; RF: 215.39 [235.62]). SVR proved to be an efficient model also for the prediction of the unpleasantness associated with olfaction (187.07 [227.32]), whereas non-significant predictions were obtained for the other algorithms (LASSO: 304.58 [219.55]; RF: 233.99 [220.96]).

Contributions to the Prediction. Having established that the data from Sharvit et al. (*18*) were most effectively modelled by a SVR under radial basis function, we identified the brain regions that contributed most to the prediction. This was assessed by running the SVR analysis on the overall population, with only one cross-validation loop for hyper-parameter optimisation. The contribution of each PCA score to the overall prediction was estimated using the method described by Guyon e al. (*40*) for non-linear Support Vector Machine kernels, which assesses the impact of the removal of each feature on the model's predictive ability. The resulting values were then back-transformed in brain space using the coefficients of the prior PCA analyses. This led to a brain map, whose values are more positive with respect to their relative importance to the prediction. Significance associated with each coordinate was obtained by re-estimating regional contributions under 10000 bootstrap resamples of the original dataset.

Results: follow-up analyses of Dilemma Events

Appropriateness ratings in both experiments were analysed using a linear mixed model with the MODALITY (thermal, olfactory) and UNPLEASANTNESS (neutral, unpleasant) of the previous cue as fixed categorical factors, and Appropriateness/Emotional Engagement of the dilemma from the validation pilot as a continuous factor. Consistent with the main analysis, data with both experiments were modelled positively with an Appropriateness effect (Exp. 1: $t_{(41.73)} = 17.59$, $p < 0.001$; Exp. 2: $t_{(52.26)} = 12.67$, $p < 0.001$), suggesting that dilemmas deemed more appropriate in the pilot were also deemed more appropriate in the main experiment. Consistently, data were negatively modulated by Emotional Engagement (Exp. 1: $t_{(51.36)} = -11.37$, $p < 0.001$; Exp. 2: $t_{(52.26)} = 12.67$, $p < 0.001$). The only other factor found to be significant was that of UNPLEASANTNESS, suggesting that individuals evaluate dilemmas as less appropriate following a cue informing about an upcoming negative event. Such effect, however, was observed exclusively in data from Experiment 1, and only when modelling dilemmas in terms of Appropriateness (Exp. 1: $t_{(222.86)} = -2.15$, $p = 0.032$; Exp. 2: $t_{(52.26)} = 1.11$, $p = 0.268$), but not in terms of Emotional Engagement (Exp. 1: $t_{(667.43)} = -1.80$, $p = 0.032$; Exp. 2: $t_{(52.26)} = 0.86$, $p = 0.388$). Furthermore, it was not found in the main analysis when DILEMMA was modelled as categorical factor. No other effect was found ($|t| \leq 1.80$, $p \geq 0.072$).

Insofar, appropriateness ratings were analysed through a linear model. However, Figure S3 shows that this measure follows a binomial distribution, with participants providing prevalently “totally appropriate” or “totally inappropriate” responses, with only few intermediate ratings. We therefore analysed the dataset in this light, by artificially categorizing participants ratings into “inappropriate” (rating < 0) and “appropriate” (rating ≥ 0) and by feeding them to a generalized linear mixed model under a binomial distribution. This analysis confirmed the previous results, by showing how participants responses were influenced by the preceding dilemma, regardless of whether this was specified as a categorical factor, or a continuous predictor of Appropriateness/Emotional Engagement (Exp. 1: $|Z| \geq 7.89$, $p < 0.001$; Exp. 2: $|Z| \geq 6.67$, $p < 0.001$). No other effect was found ($|Z| \leq 1.91$, $p \geq 0.055$).

As for the analysis of neural responses, Table S2 displays regions obtained when modelling dilemma, not as a categorical factor, but through predictors of Appropriateness and Emotional Engagement from the validation pilot. Results were almost identical to those obtained when modelling DILEMMA as a categorical factor. More specifically, in the analysis of the reading epochs, a network comprising the MPFC, TPJ, PC extending to PCC, and superior temporal sulcus was negatively modulated by Appropriateness (the least appropriate the dilemma the higher the activity) and positively modulated by Emotional Engagement. MPFC and PCC were also found in the analysis of rating epochs. In all of these analyses, the dilemma-related activity was influenced by properties of the preceding cue.

Table S1. Stimulation Events: Neural Responses. Regions displaying differential activity for the contrast HP – LP, and HD – LD (Thermal and Olfactory Events). All clusters survived correction for multiple comparisons at the cluster level (with an underlying height threshold corresponding to $p < 0.001$, uncorrected). L and R refer to the left and right hemisphere, respectively. M refers to medial activations.

	SIDE	***Coordinates***			***T***	***Cluster size***
		x	***Y***	***z***		
HP – LP						
Anterior Insula	R	36	12	6	4.61	515**
Superior Parietal Cortex	R	18	-46	68	5.42	445*
Middle Cingulate Cortex	M	0	-2	44	6.46	1657***
Supplementary Motor Area	M	-4	-14	70	5.75	
Postcentral Gyrus *(medial segment)*	M	-10	-48	70	5.37	1884***
Precuneus	M	-2	-50	54	4.92	
HD – LD						
Anterior Insula *(ventral portion)*	L	-38	16	-16	3.59	520**
Temporal Pole	L	-38	0	-22	5.65	

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ family-wise corrected for multiple comparisons at the cluster level

Table S2. Dilemma Events: Neural Responses. Regions displaying differential activity for moral – non-moral dilemmas, and showing signal linearly coupled with Appropriateness and Emotional Engagement from the validation pilot. All clusters survived correction for multiple comparisons at the cluster level (with an underlying height threshold corresponding to $p < 0.001$, uncorrected). L and R refer to the left and right hemisphere, respectively. M refers to medial activations.

	SIDE	***Coordinates***			***T***	***Cluster size***
		x	***Y***	***z***		
Moral – Non-Moral Dilemmas: Reading Epochs						
Temporo-Parietal Junction	R	60	-56	24	4.93	424**
Superior Temp. Sulcus (*anterior part*)	R	58	-14	-10	5.08	513**
Temporo-Parietal Junction	L	-52	-58	24	5.60	835***
Medial Prefrontal Cortex (*dorsal part*)	M	6	48	44	4.19	2468***
Medial Prefrontal Cortex *(rostral part)*	M	4	58	22	7.98	
Precuneus/Posterior Cingulate Cortex	M	2	-56	36	9.58	1508***
Parametric Modulation of Appropriateness *(negative effects)****: Reading Epochs***						
Temporo-Parietal Junction	R	60	-56	20	5.30	1134***
Superior Temp. Sulcus (*anterior part*)	R	58	-12	-10	6.45	1215***
Superior Frontal Gyrus	R	24	32	44	4.62	281*
Superior Frontal Gyrus	L	-18	36	44	3.72	
Temporo-Parietal Junction	L	-42	-58	16	6.01	1478***
Superior Temp. Sulcus (*anterior part*)	L	-58	-2	-16	5.71	651***
Medial Prefrontal Cortex (*dorsal part*)	M	-8	52	34	5.93	3107***
Medial Prefrontal Cortex *(rostral part)*	M	4	56	26	8.30	
Precuneus/Posterior Cingulate Cortex	M	4	-56	34	9.58	1912***
Parametric Modulation of Emotional Engagement *(positive effects)****: Reading Epochs***						
Temporo-Parietal Junction	R	60	-54	18	4.65	572**
Superior Temp. Sulcus (*anterior part*)	R	60	-12	-10	5.53	538**
Temporo-Parietal Junction	L	-42	-58	14	5.31	672***
Medial Prefrontal Cortex *(rostral part)*	M	4	56	24	7.19	2476***
Precuneus/Posterior Cingulate Cortex	M	4	-56	34	7.97	1397***
Moral – Non-Moral Dilemmas: Rating Epochs						
Medial Prefrontal Cortex (*rostral part*)	M	-4	58	8	5.13	956***
Posterior Cingulate Cortex	M	-4	-50	26	4.70	306*
Parametric Modulation of Appropriateness *(negative effects)****: Rating Epochs***						
Medial Prefrontal Cortex (*rostral part*)	M	-4	52	22	5.26	1203***
Posterior Cingulate Cortex	M	2	-52	30	5.68	813***
Parametric Modulation of Emotional Engagement *(positive effects)****: Rating Epochs***						
Medial Prefrontal Cortex (*rostral part*)	M	-2	54	22	4.97	811***
Posterior Cingulate Cortex	M	0	-52	28	4.71	295*

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ family-wise corrected for multiple comparisons at the cluster level

Table S3. Mediation analysis: rating epochs. Regions significantly modulated by the moral content of the dilemma (path *a*), significantly predicting disgust-related activity in left vAI (path *b*), and formally mediating the relationship between dilemma content and vAI (path *a*b*). Regions are highlighted if exceeding false discovery rate correction for multiple comparisons at the voxel level at $q < 0.05$. At this threshold, two coordinates in the posterior cingulate cortex are implicated in all three paths.

	SIDE	***Coordinates***			***Z***	***Cluster size***
		x	***Y***	***z***		
Dilemma → Rating Activity (path a)						
Medial Prefrontal Cortex *(rostral part)*	M	-4	62	6	8.67	8
Posterior Cingulate Cortex	M	6	-50	26	8.30	28
Rating Activity → Disgust Activity *[left vAI]* ***(path b)***						
Posterior Cingulate Cortex	M	6	-50	22	8.59	3
Dilemma → Rating Activity → Disgust Activity *[left vAI]* ***(path a*b)***						
Posterior Cingulate Cortex	M	8	-50	22	7.91	2
Conjunction: paths a ∩ b∩ a*b						
Posterior Cingulate Cortex	M	8	-50	22	7.91	2

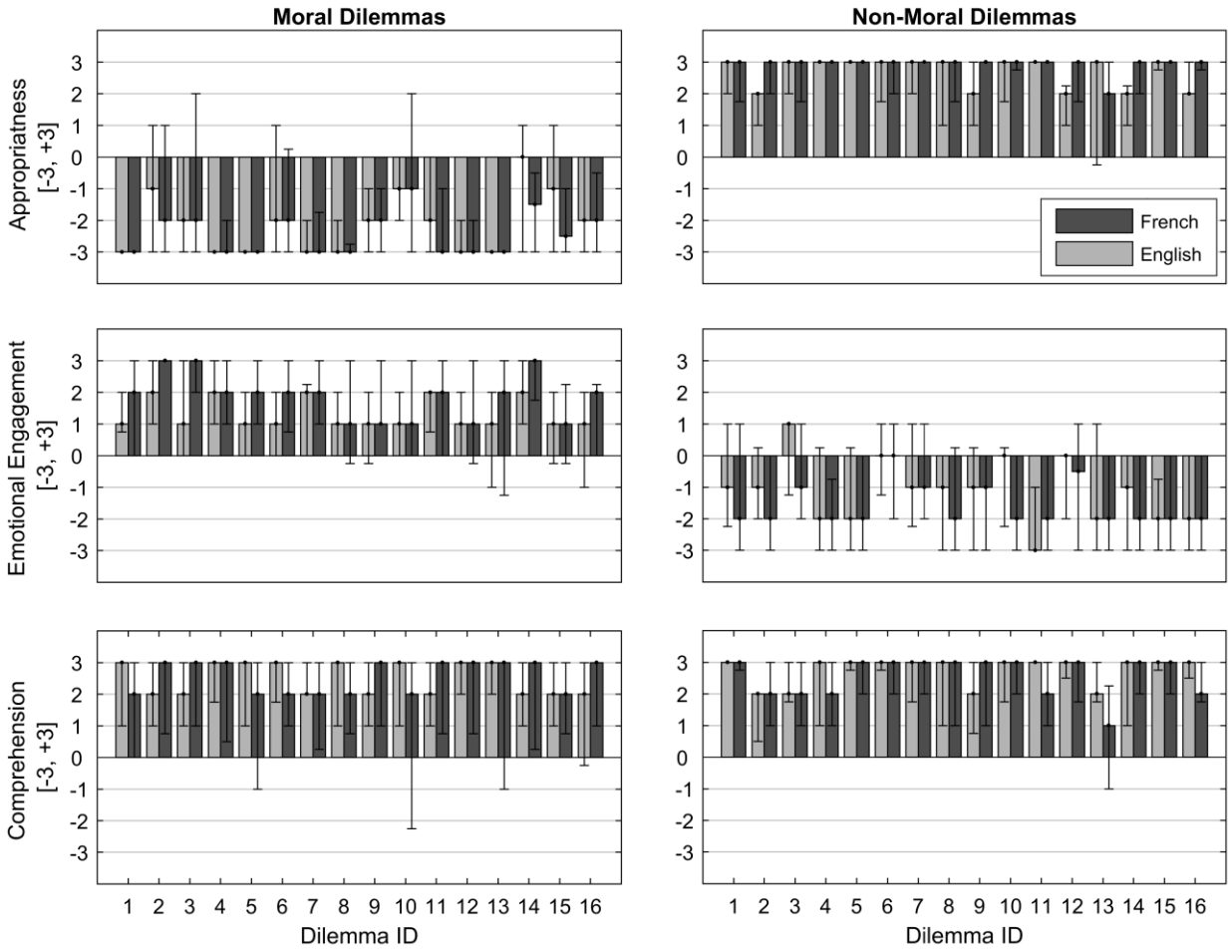

Fig. S1. Pilot data. Median ratings of appropriateness, emotional engagement, and comprehension of the 32 dilemmas that were used in the main experiment following an online pilot study. Data were obtained from 20 French and from 37 English volunteers. Ratings of the three dilemma properties ranged from -3 to +3 (see Methods section). Error bars refer to the interquartile range.

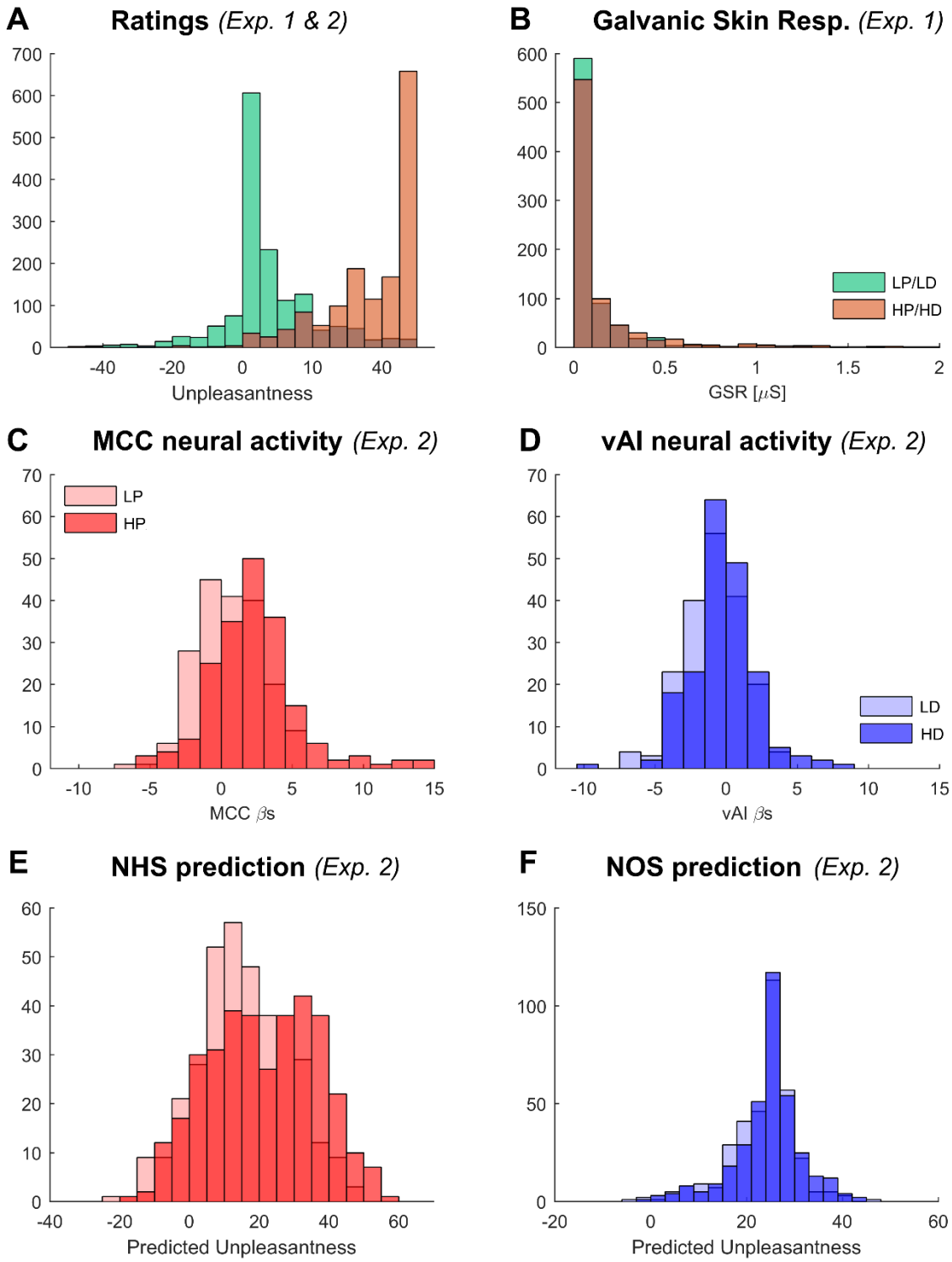

Fig. S2. Stimulation Response distribution. Histograms describing the distribution of (A) behavioural, (B) galvanic, and (C-F) neural responses to thermal and olfactory events. Histograms are colour coded according to the condition of interest. For neural responses, we display pain-related and disgust-related responses separately. *GSR*: Galvanic Skin Response; *MCC*: Middle Cingulate Cortex; vAI: ventral Anterior Insula; *NHS*: Neurological Heat Response; *NOS*: neurological Olfactory response.

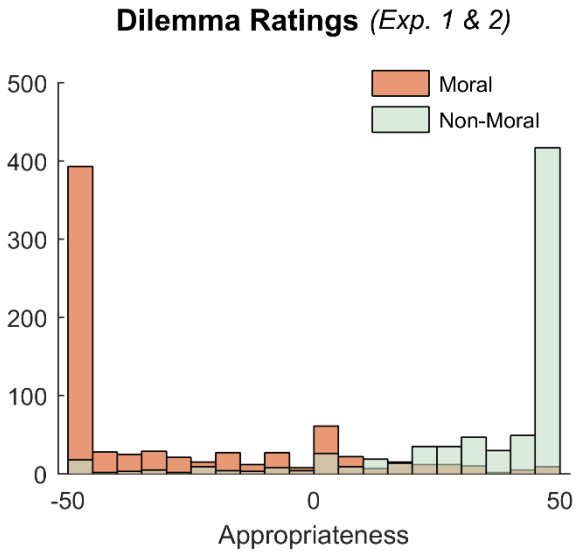

Fig. S3. Dilemma Response distribution. Histograms describing the distribution of dilemma ratings. Histograms are colour coded according to the condition of interest.