Aesthetics and morality judgments share cortical neuroarchitecture

Nora C. Heinzelmann[a1*], Susanna C. Weber[b], Philippe N. Tobler[b]

[a] *Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom*
[1] *Present address: Institute for Philosophy, Friedrich Alexander University of Erlangen-Nuremberg, Bismarckstrasse 1, 91054 Erlangen, Germany, and Munich Center for Mathematical Philosophy, Ludwig Maximilian University, Geschwister-Scholl-Platz 1, 80539 Munich, Germany*
[b] *Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Blümlisalpstrasse 10, 8006 Zurich, Switzerland, susanna.weber@econ.uzh.ch, phil.tobler@econ.uzh.ch*

* **Correspondence** should be addressed to Nora Heinzelmann (nora.heinzelmann@fau.de), Institute for Philosophy, Friedrich Alexander University of Erlangen-Nuremberg, Bismarckstrasse 1, 91054 Erlangen, Germany

Abstract

Philosophers have predominantly regarded morality and aesthetics judgments as fundamentally different. However, whether this claim is empirically founded has remained unclear. In a novel task, we measured brain activity of participants judging the aesthetic beauty of artwork or the moral goodness of actions depicted. To control for the content of judgments, participants assessed the age of the artworks and the speed of depicted actions. Univariate analyses revealed whole-brain corrected, content-controlled common activation for aesthetics and morality judgments in frontopolar, dorsomedial and ventrolateral prefrontal cortex. Temporoparietal cortex showed activation specific for morality judgments, occipital cortex for aesthetics judgments. Multivariate analyses revealed both common and distinct whole-brain corrected representations for morality and aesthetics judgments in temporoparietal and prefrontal regions. Overall, neural commonalities are more pronounced than predominant philosophical views would predict. They are compatible with minority accounts that stress commonalities between aesthetics and morality judgments, such as sentimentalism and a valuation framework.

Keywords: values, morality, aesthetics, decision-making, MVPA

1. Introduction

Beauty is in the eye of the beholder while morality is not - or so we commonly appear to believe (1, 2). Concurring with common sense, philosophy regards morality and aesthetics judgments as fundamentally different types of judgments (3-5): the former concern right or good actions, the latter beautiful or pleasant artifacts and features. In contrast, according to the sentimentalist minority tradition, *both* types of judgments are expressions of approbation or disapprobation (6-10). For example, some (11) argue that value judgments are constituted by empirically observable embodied emotions.

At a first glance, neuroscientific evidence appears to support both views, as it reports common as well as specific activations for morality and aesthetics judgments. Common activations arise in the orbitofrontal cortex (OFC) and the adjacent medial prefrontal cortex and frontal pole (12-17). Morality judgments have been linked specifically to the occipital cortex, middle frontal gyrus, temporoparietal junction, posterior and anterior cingulate cortex, middle temporal gyrus, and precuneus (12, 13, 18, 19). Aesthetics judgments have been associated with (different parts of) the occipital cortex, putamen, and OFC (12, 20, 21).

However, a closer look reveals limitations of this literature. First, few studies have investigated both morality and aesthetics judgments within the same participants (12-14) and one of them has not directly compared the two judgment types (14). Second, all studies involved different stimulus material and differences in visual processing for assessing the two judgment types, which makes them vulnerable to confound (cf. 22's critique of 21). For instance, one study (13) used pencil drawings for morality judgments and photographs for aesthetics judgments. Any difference in neural activity between the two tasks may thus be attributed to differences between drawings and photographs, not between morality and aesthetics judgments. Third, although some studies did use a (single) control task, it concerned yet another set of stimuli (13, 14). Presumed commonalities of morality and aesthetics judgments may thus actually have been induced by the control stimuli or control task that served as a common reference for the analysis.

Our study is a tailored effort to inform the debate surrounding moral and aesthetic value judgments, drawing on meta-ethics and meta-aesthetics for operationalization and using tight experimental control. Paradigmatic morality judgments ascribe a moral property like moral goodness to an action (23). Exemplary aesthetics judgments concern the beauty of an object of art (24). Accordingly, we asked participants to judge the beauty of artistic images and the moral goodness of the actions depicted in the images (Figure 1a). One common stimulus set (see Supplemental Information) served as the basis for both morality and aesthetics judgments as well as for two respective control judgments: rating the speed of the action served as a control task to rating its moral goodness, and rating the age of the artwork as a control for rating its beauty. This ensured that commonalities of morality and aesthetics judgments could not be due to peculiarities of the control task and its stimulus material. We also collected eye-tracking data to account for potential differences in visual processing.

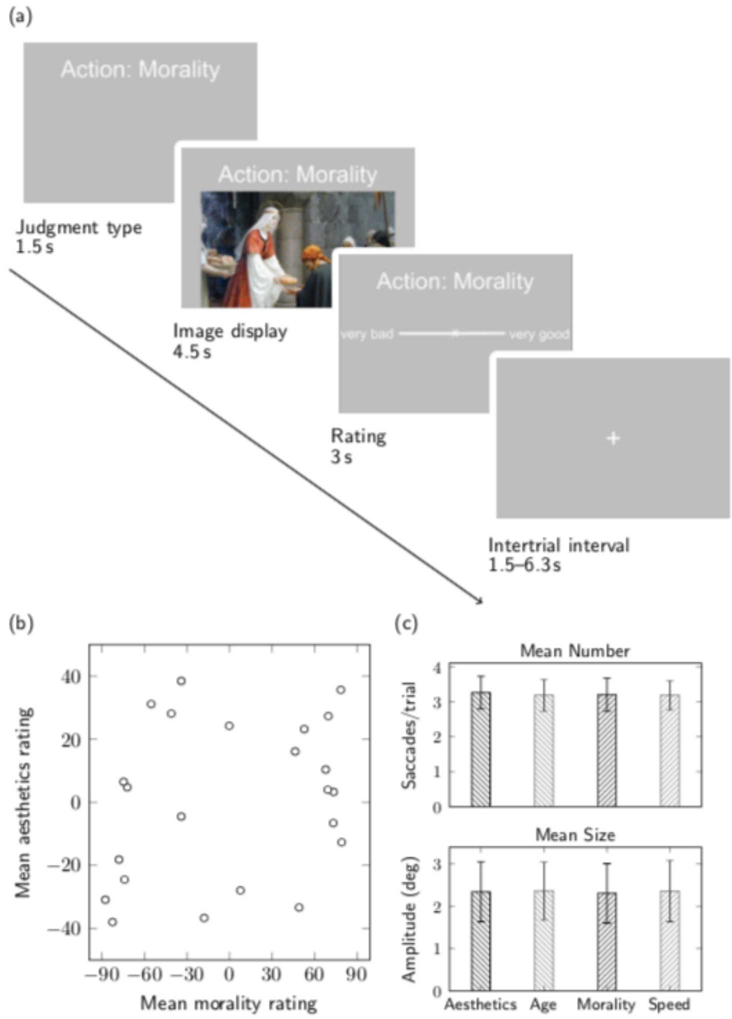

Figure 1 Task and behavior. (a) The beginning of each trial specified judgment content and type. After 1.5s, a randomly selected image appeared below the specification. Participants had 4.5s to assess the image. They indicated and confirmed their judgment on a continuous rating scale (3s max). The orientation of the scale and the initial position of the cursor on the scale were randomized across trials, to prevent motor preparation during the image display phase. (b) Image ratings. The mean morality and aesthetics ratings for each of the 24 images were not correlated with one another. (c) Mean saccade number (top) and size (amplitude; bottom) during *image display* for each condition, averaged across images and participants. Error bars indicate standard deviations. There was no significant difference between the four judgment types for both saccade number and size.

As it highlights commonalities of the two judgment types, a sentimentalist hypothesis would predict common activations for morality and aesthetics judgments. Conversely, the majority view would expect different regions to be involved in each of the two types of judgment, or at least different representations within the same region. To assess the possibility of such fine-grained judgment-specific activity patterns within a given region, we used multivariate methods. Our study thus aims to inform the debate between sentimentalist and standard views by overcoming methodological concerns affecting the current literature, and to advance empirical research on sentimentalism and value judgments more generally.

2. Results

2.1 Behavioral results

First, we assessed whether morality and aesthetics judgments were correlated (Figure 1b). We found no significant relation across images ($r=0.33$, $p=0.12$) or participants ($r=0.19$, $p=0.43$). Thus, when making the two kinds of judgments, participants were engaged in two differentiable activities. Moreover, knowing the aesthetic status of a given image provided little information about the moral status of the action it depicted (and vice versa), even though both types of judgments concerned the same stimulus material.

Second, both the number and size of saccades were similar for all conditions ($F(3,112)=0.03$, $p=0.99$ for number, $F(3,108)=0.04$, $p=0.99$ for size; Figure 1c). There were no significant differences between morality and aesthetics conditions (size: $t=-1.60$, $p=0.12$; number: $t=-1.42$, $p=0.17$). Pupil size also did not differ significantly between the four conditions ($F(3,116)=0.03$, $p=0.99$). Thus, participants appeared to employ similar visuo-motor processes for all types of judgments.

2.2 Neuroimaging results

2.2.1 Common univariate activations for morality and aesthetics judgments

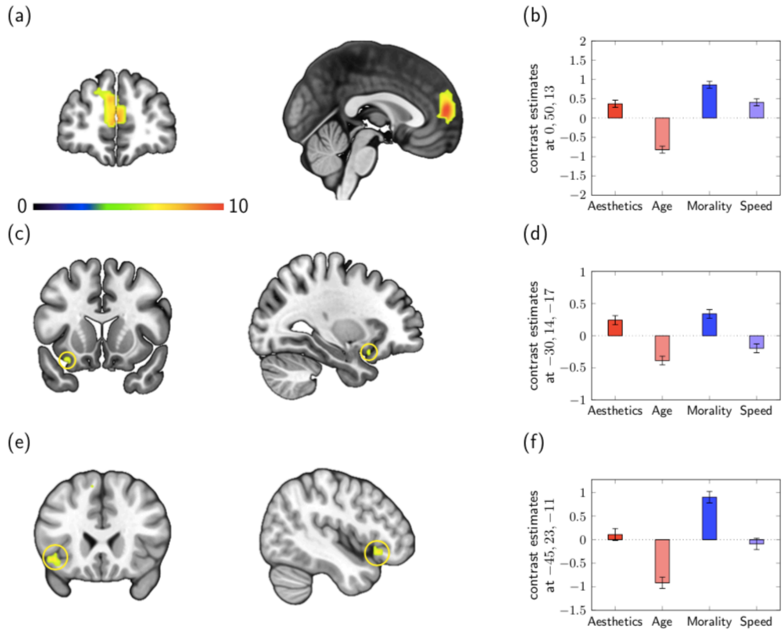

Figure 2 Common univariate activations for morality and aesthetics judgments. Brain sections show regions activated for both the contrast (morality-speed) and the contrast (aesthetics-age), both whole-brain FWE corrected, p<0.05 and inclusively masked, thus representing above-threshold voxels for both contrasts. (a) Common cluster in frontal pole (extending into anterior cingulate cortex). (b) Contrast estimates at peak (0,50,13). (c) Common cluster in insula. (d) Contrast estimates at peak (-30,14,-17). (e) Common cluster in orbitofrontal cortex. (f) Contrast estimates at peak (-45,23,-11). All coordinates are in MNI space. Color bars indicate t-values, error bars standard error.

To test the sentimentalist hypothesis, we determined whether morality and aesthetics judgments involve the same brain regions. Separate analyses (both $p<0.05$, FWE-whole brain corrected; see Supplementary Table 1) showed similar activations for the two content-controlled contrasts (morality-speed) and (aesthetics-age). This impression was confirmed by inclusive masking, which identified common activations primarily in regions of the prefrontal cortex, such as the frontal pole/anterior cingulate (Figure 2a), dorsomedial prefrontal (Figure 2c), and ventrolateral prefrontal cortex (Figure 2e). All these regions showed stronger activation for morality and aesthetics judgments than for their respective controls (Figures 2b, d, f; Supplementary Table 2 lists all activations). Thus, morality and aesthetics judgments commonly activate regions in medial and lateral prefrontal cortex. Importantly, these common activations cannot be explained by differences in the stimulus material of the control task or by the use of one common control task for both morality and aesthetics judgments.

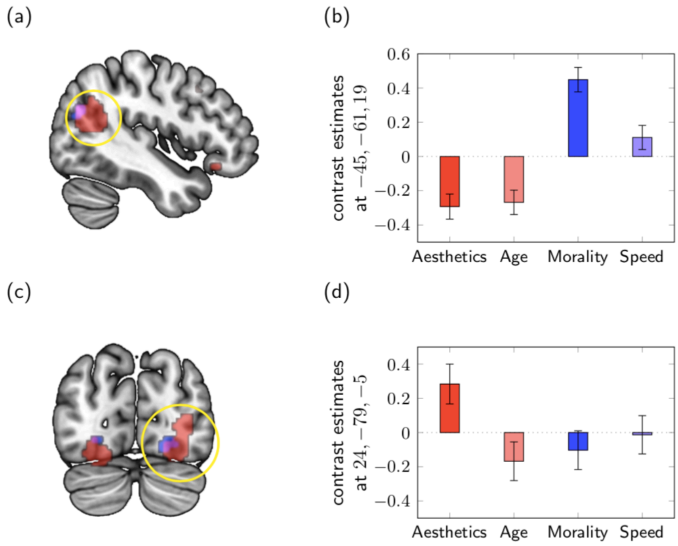

Figure 3 Specific univariate activation. Areas of activations for content-uncontrolled differences between morality and aesthetics judgments are shown in red, content-controlled differences in blue. Overlap is shown in pink. (a) Specificity for morality judgments. Regions activated for the contrast (morality-aesthetics) are shown in red and for (morality-speed)-(aesthetics-age) in blue. (b) Contrast estimates at overlap (-45,-61,19). (c) Specificity for aesthetics judgments. Regions activated for (aesthetics-morality) are shown in red and for (aesthetics-age)-(morality-speed) in blue. (d) Contrast estimates at overlap (24,-79,-5). All coordinates are in MNI space. Error bars indicate standard error of the mean.

2.2.2 Specific univariate activations for morality judgments

To localize activations specifically related to morality or aesthetics judgments, we used a two-stage approach (see Methods). The first-stage contrast (morality-aesthetics) identified tentative activation differences between the two judgments of interest without controlling for the differences in the content of judgment (action vs. image). The second-stage contrast [(morality-speed)-(aesthetics-age)] then interrogated the identified regions whilst controlling for content. For morality judgments, the content-uncontrolled first-stage contrast identified stronger activity for morality than for aesthetics judgments primarily in the medial and lateral prefrontal cortex and temporoparietal junction (TPJ; whole-brain corrected; Supplementary Table 3 lists all activations). The second stage revealed that TPJ activity was indeed morality-specific as indicated by the content-controlled contrast at a whole-brain corrected threshold.

To independently assess these judgment-specific findings and apply our tightly controlled approach to existing research, we performed ROI analyses.

Specifically, we drew on a meta-analysis of previous univariate tests that consistently found activity related to morality (25, Figure 4a). Again, we applied our two-stage approach. The contrast (morality-aesthetics) revealed TPJ and dorsomedial prefrontal cortex ($p<0.05$, FWE corrected for all voxels in the combined ROI; Supplementary Table 5). Second, we controlled for the content of judgment, using the contrast [(morality-speed)-(aesthetics-age)]. The TPJ, but not the prefrontal cluster, again showed significant activity at $p<0.05$, FWE small-volume corrected (Figure 4b, Supplementary Table 5). Thus, while the TPJ appears to preferentially encode morality judgments, the prefrontal cluster identified by the same first-stage contrast instead seems to reflect differences between contents (actions vs. images) rather than differences between kinds of judgments (morality vs. aesthetics). The ROI-based findings in TPJ thus converge with the whole-brain evidence for morality-specific activity in the TPJ mentioned above.

2.2.3 Specific univariate activations for aesthetics judgments

At the whole-brain level, we aimed to identify regions specific for aesthetics judgments and again used the two-stage approach. First, the contrast (aesthetics-morality) revealed activation in orbitofrontal, lateral prefrontal, and occipital cortices and in the cerebellum at the threshold of $p<0.05$, whole-brain FWE corrected. (Supplementary Table 3, Figure 3c red). The second stage contrast [(aesthetics-age)-(morality-speed)] revealed that the occipital activity but not the activity in the other regions was indeed specific for aesthetics judgments, once we controlled for judgment content (Figure 3c, pink overlaps; see Supplementary Information for the specificity analysis of control judgments).

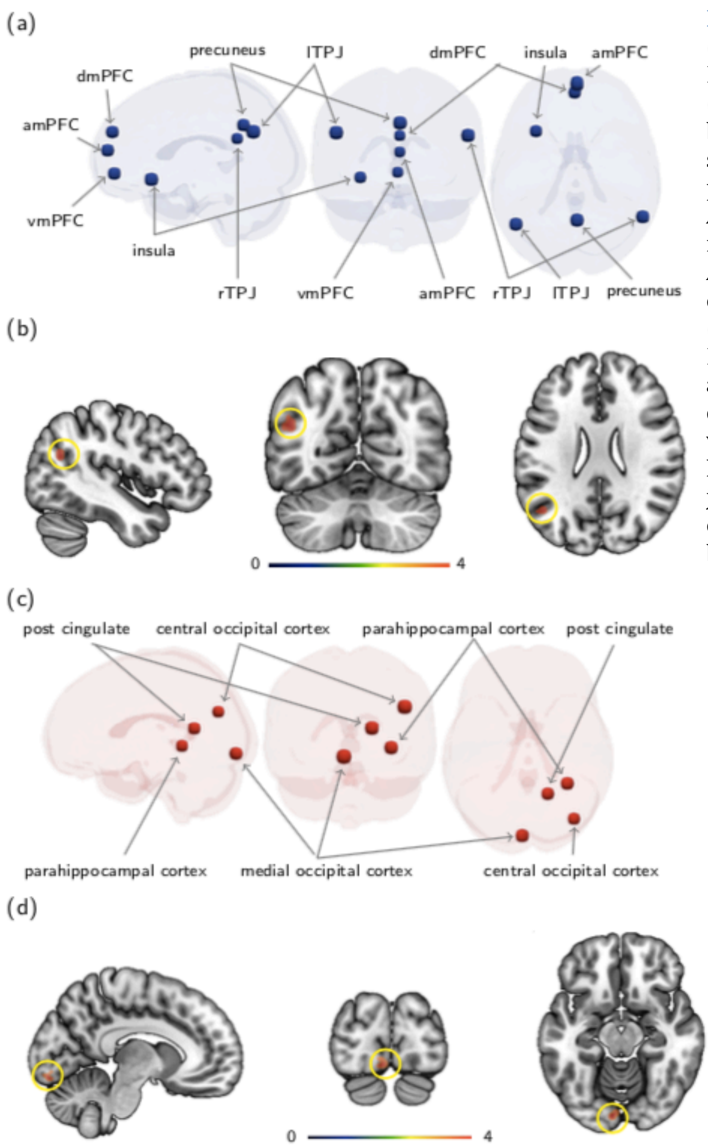

Figure 4 Region of interest (ROI) analyses. (a) Morality ROI based on a meta-analysis (25). (b) Activations revealed by the contrast [(morality-speed)-(aesthetics-age)], $p<0.05$, FWE corrected. (c) Aesthetics ROI based on a meta-analysis (26). (d) Activations revealed by the contrast [(aesthetics-age)-(morality-speed)], $p<0.05$, FWE corrected. Abbreviations: am, anterior medial; dm, dorsomedial; vm, ventromedial; lTPJ/rTPJ, left/right temporoparietal junction; PFC, prefrontal cortex; post, posterior. Color bars represent t-values.

We also performed an ROI-analysis for aesthetics analogously to the one for morality judgments. We relied on previous studies using univariate analyses that consistently found activity related to aesthetics judgments (26, Figure 4c). Using our two-stage approach, we first identified two clusters in the right fusiform gyrus ROI and one in the middle occipitotemporal gyrus ROI for the contrast (aesthetics-morality) at a threshold of $p<0.05$, FWE corrected for all voxels in the combined ROI. Second, we controlled for judgment content within the combined aesthetics ROI with the contrast [(aesthetics-age)-(morality-speed)], applying a threshold of $p<0.05$, FWE corrected for all the voxels in the combined ROI. We found activation in the middle occipital cortex (Figure 4d, Supplementary Table 5). Thus the fusiform clusters from the first stage seem to have arisen from content of judgment (images rather than actions) and not specifically from making an aesthetics rather than a morality judgment. The ROI-based findings in occipital cortex converge with the whole-brain evidence for aesthetics-specific activity in that region.

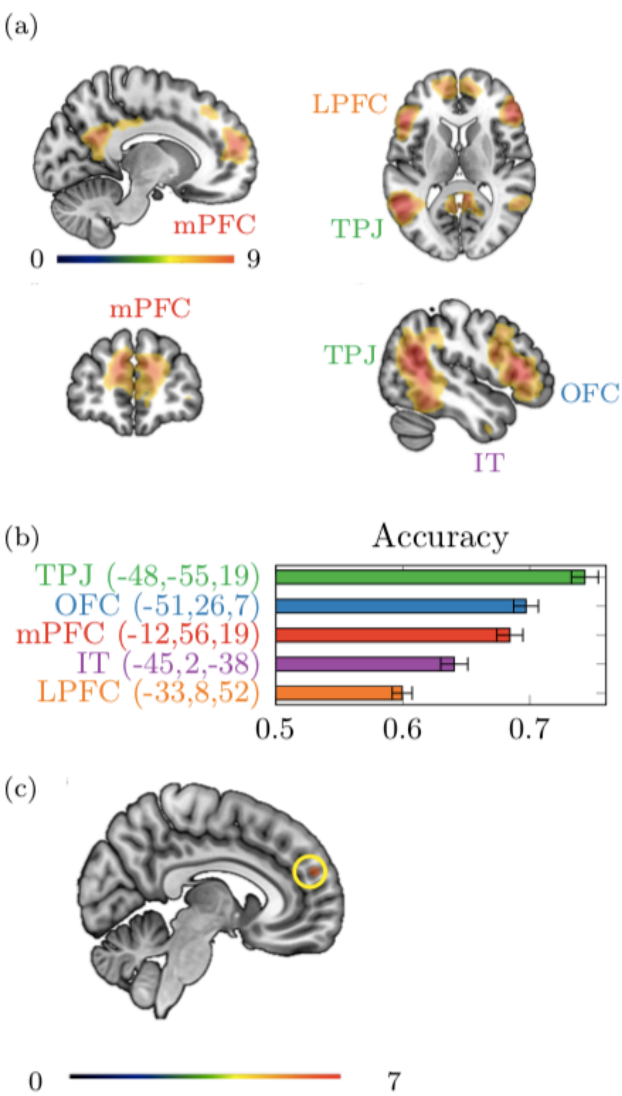

Figure 5 Common multivariate representations for morality and aesthetics judgments as revealed by cross-classification analyses. (a) In a first analysis, we trained a classifier to distinguish between morality and speed and tested it on aesthetics vs. age, *and vice versa*. All cross-classifications are significant at $p<0.05$, voxel-wise whole-brain FWE corrected (see Supplementary Table 6 for a full list). Medial and lateral prefrontal cortex (mPFC, LPFC), left temporoparietal junction (TPJ), orbitofrontal cortex (OFC), and inferior temporal cortex (IT) show common representations of judgment types. (b) Test accuracies in peak voxels of the identified regions (chance level: 0.5). Error bars represent standard error, color bars t-values. (c) A second cross-classification analysis revealed common multivariate representation in medial PFC for morality and aesthetics judgments. Here we trained a classifier to distinguish between high and low morality ratings and tested it on high versus low aesthetics ratings, and *vice versa*. Cross-classification is significant at $p<0.05$, voxel-wise whole-brain FWE corrected. The figure shows the peak at MNI coordinates 6, 53, 25. This region overlapped with the medial prefrontal region identified in (a).

2.2.4 Multivariate representations of morality and aesthetics judgments

To investigate how morality and aesthetics judgments are represented on a finer level (27), we performed multivoxel pattern analyses (MVPAs). As for the univariate analysis, we proceeded in two steps. First, we investigated common representations of the two judgment types. To do so, we performed a cross-classification MVPA which identified regions where patterns of activity distinguished both morality from speed judgments and aesthetics from age judgments. Specifically, we trained a classifier to distinguish morality versus speed judgments, and tested it on aesthetics versus age judgments, and *vice versa*. Cross-classification indicates that information distinguishing one type of value judgment (say, morality judgments) from its content-controlled counterpart (speed) is represented similarly as information distinguishing the other type of value judgment (aesthetics) from its content-controlled counterpart (age). We found cross-classification to be significantly above chance in medial and lateral prefrontal cortex, orbitofrontal cortex, TPJ and inferior temporal cortex at a $p<0.05$ whole-brain FWE corrected threshold (Figure 5a; see Supplementary Table 6 for a full list). Accuracy estimates at peak voxels indicate that cross-classification was most accurate in TPJ, followed by OFC and mPFC (Figure 5b). These findings suggest that prefrontal and temporoparietal regions represent morality and aesthetics judgments similarly, but differently from factual judgments about the same content.

In a second cross-classification MVPA, we assessed commonality between morality and aesthetics judgments without using the control judgments. Specifically, we trained a classifier to distinguish high from low morality ratings and tested it on high versus low aesthetics ratings, and *vice versa*. We found information distinguishing high from low ratings in both value domains to be represented in the medial prefrontal cortex at a $p<0.05$ whole-brain FWE corrected threshold (Figure 5c). Thus, the two independent cross-classification analyses both suggest that the medial prefrontal cortex provides a common neural basis for morality and aesthetics judgments.

To test for judgment-specific multivariate representations, we performed a whole-brain searchlight MVPA by training a classifier to distinguish between aesthetics and morality judgments. The TPJ and lateral prefrontal cortex differentially represented the two judgments (shown in red in Figure 6a; $p<0.05$, whole-brain FWE corrected, see Supplementary Table 7 for a full list). These regions partially overlapped with those identified in the cross-classification first-step analysis (shown in blue in Figure 6a, overlap in pink). Again, accuracy estimates at peak voxels for the morality versus aesthetics classifier indicate that decoding was most accurate for TPJ (Figure 6b). Together, these data suggest that value-sensitive regions in temporoparietal and lateral prefrontal cortex also differentially represent morality and aesthetics judgments.

3. Discussion

Our study is a tailored effort to determine whether morality and aesthetics judgments differ in their neural architecture or not. It has two main findings: first, the two judgment types share a common, large-scale architecture primarily in anterior and medial prefrontal cortex. Second, we find judgment-specific activations and multivariate representations in occipital, temporoparietal, and lateral prefrontal regions.

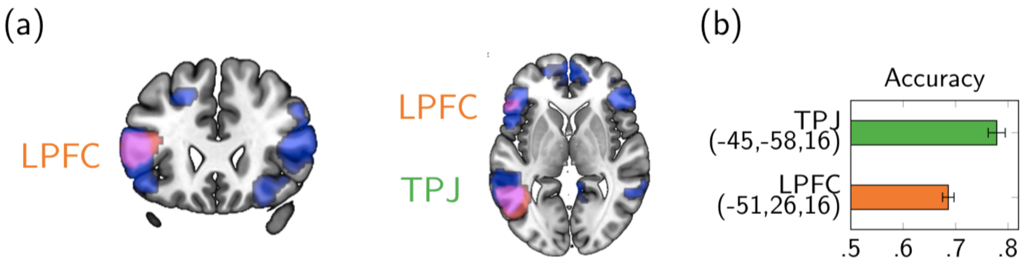

Figure 6 Specific multivariate representations for morality and aesthetics judgments. We trained a classifier to distinguish activity patterns for morality vs. aesthetics. The identified regions carried significant information about judgment type at p<0.05, voxel-wise whole-brain FWE corrected (see Supplementary Table 7 for a full list). (a) Activity patterns yielded by the morality vs. aesthetics classifier (red) identified specific multivariate representations in frontal and temporal regions. To assess whether these regions involved regions showing common multivariate representations, we replotted the data from the cross-classification analysis described in Figure 5a (blue). Overlap in temporoparietal junction (TPJ) and lateral prefrontal cortex (LPFC) shown in pink. (b) Test accuracies of the morality versus aesthetics classifier in peak voxels in TPJ and LPFC (chance level: 0.5). Error bars represent standard error.

Our univariate analyses found less specific activation for both morality and aesthetics judgments than previous reports (12-14, 28). This difference can be explained by the increased experimental control of our study: first, we employed the same stimulus material for the two judgment types and thereby excluded differences in visual input as potential explanation for differences. Second, we used control judgments whose contents were matched to those of the value judgments to account for the fact that morality judgments concern actions whereas aesthetics judgments concern images. In hindsight, our data justify this approach: without the use of control judgments, our study would have falsely inferred strong univariate differences between the two types of judgments, e.g., in temporal cortex (Supplementary Table 3). However, once we controlled for the differences in judged objects, many of these alleged differences between morality and aesthetics judgments disappeared.

Judgments about morality and aesthetics are both value-based whereas judgments about age and speed are not. Accordingly, our findings of common activation in prefrontal regions associated with value processing are in-keeping with the view that value plays a role for both judgment types. By extension, our study speaks to the ongoing debate about whether value is encoded in a domain-specific or domain-general fashion. If the brain is to compare and decide between vastly different rewards, then it requires a domain-general common currency signal (29, 30-32). On the other hand, adaptive and model-based behavior may require keeping track of the value of specific goods. Within the gustatory domain, a common currency signal has been identified in medial PFC whereas OFC has been associated with goods-specific value signals (33, 34). Our work proposes a similar scheme with regard to value-based judgment types: we find domain-general activation in the medial and ventrolateral PFC and, in MVPA analyses, domain-specific activation in central OFC. Interestingly, we do not find activation in the vmPFC which has been identified as a 'common currency' region for value in previous studies (e.g., 35). This might be due to a different operationalisation of 'value': the detection of aesthetic beauty and moral goodness in our study is at best indirectly beneficial to our participants whilst monetary payoffs or social status are more obviously advantageous to them and associated with activation in the vmPFC. In line with this notion, self-regarding values are correlated with more ventral regions of the mPFC, other-regarding values with more dorsal ones (36, 37).

Within the debate about whether morality and aesthetics judgments differ fundamentally or not, sentimentalism (11) states that both are expressions of approval or disapproval. Accordingly, sentimentalism would predict common brain activation for morality and aesthetics judgments. Our test of this prediction yields one positive and one negative result. On the positive side, our univariate and multivariate data (Tables S2 and S6) concur with the sentimentalist prediction in that both value judgments share common neuroarchitecture. Thereby, they contest rival approaches that stress essential differences between morality and aesthetics judgments, such as the Kantian view according to which aesthetics judgments are based on disinterested pleasure whereas morality judgments are based on reason (4, 5).

On the negative side, our findings challenge sentimentalism as well. First, we find specific representations of different value judgments in more posterior regions such as TPJ for morality and occipital cortex for aesthetics (Supplementary Table 3). Second, at a finer-grained neural level, our multivariate and ROI analyses (Figure 3, Tables S5 and S7) confirmed judgment-specific processes in temporoparietal, occipital, and lateral prefrontal.

Note that our study does not make any quantitative claim about sentimentalism and its rivals from the degree of commonality between aesthetics and morality judgments. Instead, we show categorically that there are both commonalities and differences between the two types of judgments.

In conclusion, then, our data indicate that morality and aesthetics judgments share particularly medial prefrontal neuroarchitecture, possibly due to the fact that both of these judgments are based on value. Yet, contrary to sentimentalism, particularly lateral prefrontal and temporoparietal regions do not conflate the two judgment types, indicating that the brain differentiates the morally good from the aesthetically pleasing.

4. Materials and methods

We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Study data, digital study materials, and analyses codes are available on OSF (https://osf.io/9q286), OpenNeuro (https://openneuro.org/datasets/ds002732) and by request, subject to copyright constraints. Some of our stimuli contain details of artwork that may be subject to copyright constraints. We are happy to share stimuli provided we are not prevented to do so by legal restrictions. No part of the study procedures or analyses was pre-registered prior to the research being conducted.

4.1 Participants

30 healthy, right-handed participants (19 females, mean age: 23) participated in the study. We chose the sample size a priori on the basis of previous research (12-14, 20) with sample sizes ranging from 10 to 28. Data from three participants (2 males, 1 female) were excluded from the analysis because two of them showed excessive head movements and the third failed to respond in 84% of all trials. The study was approved by the ethics committee of the Canton of Zurich.

4.2 Task

Before receiving specific instructions, participants viewed all 24 images outside the scanner and described the depicted action to the experimenter. This prevented misunderstandings about what was shown and allowed participants to quickly identify the depicted actions in the main experiment. In the scanner, participants completed three sessions of about 20 minutes each (Figure 1a). During each session, they provided four ratings for each of the 24 images; these concerned (i) the moral goodness or badness of the action depicted, (ii) the speed of this action (non-moral control judgment), (iii) the aesthetic beauty or ugliness of the image, and (iv) its age (non-aesthetic control judgment). All trials were randomized within each session but spread according to repetition constraints (e.g., no image appeared twice consecutively). Subject to copyright restrictions, images are available upon request and online (see Methods).

Trials lasted 12.1s on average. At the beginning of each trial, participants were informed about the judgment type. After 1.5s, an image appeared and stayed on the screen for 4.5s. Participants had been instructed to make a judgment during this phase of the trial based on their personal opinion (e.g., 'an 18th century painting might seem old to you, or new'). The subsequent phase allowed just enough time (3s) to enter the rating using a continuous scale. It ranged from 'very good' to 'very bad' in the morality condition, from 'very beautiful' to 'very ugly' in the aesthetics condition, from 'very fast' to 'very slow' in the speed condition, and from 'very old' to 'very new' in the age condition. The orientation of the rating scale was randomized for each trial. Using a trackball, participants moved the cursor to the position on the scale that reflected their rating, and confirmed their choice with a button press. After a variable inter-trial interval (mean: 3.1s), the next trial started.

4.3 fMRI data analysis

We fitted a parametric general linear model (GLM) to the data of each subject (see Supplemental Information for details about fMRI data acquisition and preprocessing). Four regressors specified the onset for modeled morality, aesthetics, speed, or age judgments. The duration of the regressors corresponded to the sum of the 4.5s the image was shown and the response time in the given trial. Additional regressors of no interest were the ratings given on each trial (modeled as parametric modulators; Supplementary Table 4), the onsets of missed trials, motion regressors, and eye tracking data to account for oculomotor activity and potential differences in visuo-motor processing. We used SPM's canonical hemodynamic response function (HRF) to convolve the task-related regressors and the default high-pass filter of 128s.

To identify regions of common activation induced by morality and aesthetics judgments, we used the onset regressors in each participant to form the contrasts (morality-speed) and (aesthetics-age). Inclusive masking of these two contrasts at the group level then identified regions of common activation for morality and aesthetics judgments. Similar to a conjunction null analysis, inclusive masking determines the intersection of the above-threshold voxels for the morality-speed contrast and the above-threshold voxels for the aesthetics-age contrast. To test for areas of judgment-specific activation, we used a two-stage approach.

At the first stage, we formed the single-subject-level contrasts (morality-aesthetics) and (aesthetics-morality). A significant difference in these contrasts is necessary but not sufficient to identify judgment-specific activations because these contrasts concerned different content (actions for morality judgments vs. images for aesthetics judgments). To qualify the findings from the first stage, we therefore also formed the content-controlled contrasts [(morality-speed)-(aesthetics-age)] and [(aesthetics-age)-(morality-speed)] at the individual-subject level. Differential activity in these contrasts cannot be due to differences in content because speed and age judgments had the same respective contents. The second stage thus allowed us to investigate whether activations identified in the first stage are indeed due to differences between the two value judgments or due to differences in the content of these judgments. In all cases, we used the resulting individual contrast images for random effect analysis at the group level. All results reported are from statistic images assessed for peak-level significance, $p<0.05$ FWE whole-brain corrected.

We conducted two regions of interest (ROI) analyses for morality and aesthetics judgments, respectively. To create the ROIs, we relied on two meta-analyses from the literature (25, 26) and used the *WFU Pickatlas* extension in SPM (38, 39). In each case, we formed a combined ROI by creating a sphere (r=6mm) around peak voxels of clusters reported. For the morality ROI analysis, we relied on results for explicit, affective, cognitive, or other-assessments of morality, averaged for overlapping clusters (25). For the aesthetics ROI, we used the peak voxels of clusters reported for aesthetics evaluation (26).

We conducted three MVPAs to test for finer-grained neural representations common and specific for morality and aesthetics judgments using *The Decoding*

Toolbox (40). To test for content-controlled commonalities, we performed a two-way cross-classification MVPA. We fitted the same GLM that we used for the univariate analysis to the unsmoothed functional images of each participant. We then trained an L2-norm support vector machine (SVM; 41, 42) in a cross-validated leave-one-run-out searchlight decoding analysis. Specifically, we trained this classifier to distinguish morality vs. speed judgments and tested it on aesthetics vs. age judgments, *and vice versa*. At the group level, we performed t-tests using SPM on smoothed, individual accuracy maps (smoothing kernel: 6 mm FWHM). We applied an FWE whole-brain corrected, $p<0.05$ threshold. Successful cross-classification implies common differences between experimental and control judgments (Figure 5 and Figure 6, blue). Using the CosmoMVPA package (43), we also replicated these findings in a permutation analysis by Stelzer and colleagues (44) that stringently controls for false positives (see Supplementary Information for details and results).

Second, we performed an additional cross-classification MVPA to identify common representations for morality and aesthetics independent of control judgments. For each participant, we first fitted a factorial model to the un-smoothed images using a median split of moral and aesthetic ratings. We then trained an L2-norm SVM to distinguish high versus low morality ratings and tested it on high versus low aesthetics ratings, and *vice versa*. Here too, we applied FWE whole-brain correction, $p<0.05$ (Figure 5c).

Third, to assess judgment-specific representations, we performed a cross-validated leave-one-run-out searchlight decoding analysis with a default searchlight radius of 4 voxels. Specifically, we used beta images from our GLM to train an L2-norm SVM as a classifier aiming to distinguish between aesthetics and morality judgments. We again applied an FWE whole-brain corrected, $p<0.05$ threshold. Resulting representations reflect neural differences between morality and aesthetics judgments (Figure 6, blue). Again, we verified our findings using a permutation approach.

Author contributions

Nora Heinzelmann: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition; **Philippe Tobler**: Conceptualization, Methodology, Software, Validation, Formal Analysis, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition; **Susanna Weber**: Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization

Declaration of interest

None

Acknowledgments

This study was supported by the Swiss National Science Foundation (Grants PP00P1_128574, PP00P1_150739, and 100014_165884), the University of Cambridge (Fieldwork Funds), the UK Arts and Humanities Research Council (Grant 04386), and the Faculty of Philosophy, Ludwig Maximilian University of

Munich. We acknowledge the Neuroscience Center Zurich and thank Guy Kahane and Thorsten Kahnt for helpful discussions on study design, Karl Treiber for scanning support, Adrian Etter for eyetracking expertise, Chris Burke, Justin Chumbley, Björn Lindström, Nick Oosterhof, Alexander Soutschek and Anna Stöckl for input on data analyses, and Bahador Bahrami, Azade Doğan and Ophelia Deroy for comments on the manuscript. We also acknowledge feedback from audiences at conferences for Experimental Philosophy (Nottingham, 2015), the European Society for Philosophy and Psychology (Tartu, 2015), the European Society for Aesthetics (Barcelona, 2016), and on the *Social Brain* (Aegina, Greece, 2017).

References

1. Goodwin, G. & Darley, J. The psychology of meta-ethics. *Cognition* **106,** 1339-1366 (2008).

2. Hanson, L. Moral realism, aesthetic realism, and the asymmetry claim. *Ethics* **129,** 39-69 (2018).

3. Came, D. Moral and aesthetic judgments reconsidered. *J Value Inquiry* **46,** 159-171 (2012).

4. Strandberg, C. A structural disanalogy between aesthetic and ethical value judgements. *Brit J Aesthet* **51** 51-67 (2011).

5. Kant, I. *Critique of Judgment*, transl Meredith, J. (Oxford University Press, Oxford, 1790/1928).

6. Hume, D. *An Enquiry concerning the Principles of Morals*, ed Beauchamp, T. (Oxford University Press, Oxford/New York, 1751/1998).

7. Hume, D. *Of the Standard of Taste*, ed Miller, E. (Liberty Fund, Indianapolis, 1757/1985).

8. Hutcheson, F. *An Inquiry Into the Original of Our Ideas of Beauty and Virtue in Two Treatises*, ed Leidhold, W. (Liberty Fund, Indianapolis, 1725/2004).

9. Hutcheson, F. *An Essay on the Nature and Conduct of Passions, With Illustrations on the Moral Sense*, ed Garrett, A. (Liberty Fund, Indianapolis, 1728/2002).

10. Shaftesbury, Third Earl of. *An Inquiry Into Virtue and Merit. Characteristicks of Men, Manners, Opinions, Times, Vol.2*, ed Uyl, D. (Liberty Fund, Indianapolis, 1714/2001).

11. Prinz, J. *The emotional construction of morals* (Oxford University Press, New York, 2007).

12. Wang, T. et al. Is moral beauty different from facial beauty? Evidence from an fMRI study. *Soc Cogn Affect Neur* **10**, 814-823 (2015).

13. Avram, M. et al. (2013) Neurofunctional correlates of esthetic and moral judgments. *Neurosci Lett* **543,** 128-132 (2013).

14. Tsukiura, T. & Cabeza, R. Shared brain activity for aesthetic and moral judgments: implications for the Beauty-is-Good stereotype. *Soc Cogn Affect Neur* **6**, 138-148 (2011).

15. Jacobsen, T., Schubotz, R., Höffel, L. & Von Cramon, D. Brain correlates of aesthetic judgment of beauty. *NeuroImage* **29,** 276-285 (2006).

16. Ishizu, T. & Zeki, S. Toward a brain-based theory of beauty. *PLoS One* **6,** e21852 (2011).

17. Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F. & Grafman, J. The neural basis of human moral cognition. *Nat Rev Neurosci* **6,** 799-809 (2005).

18. Kahane, G. et al. The neural basis of intuitive and counterintuitive moral judgment. *Soc Cogn Affect Neur* **7,** 393-402 (2011).

19. Bzdok, D. et al. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct Funct* **217,** 783-796 (2012).

20. Kawabata, H. & Zeki, S. Neural correlates of beauty. *J Neurophysiol* **91,** 1699-1705 (2004).

21. Zhang, W. et al. Neural substrates of embodied natural beauty and social endowed beauty: an fMRI study. *Sci Rep-UK* **7,** 7125 (2017).

22. Nadal, M., Gallardo, V. & Marty, G. Commentary: Neural substrates of embodied natural beauty and social endowed beauty: an fMRI study. *Front Hum Neurosci* **11,** 1-2 (2017).

23. Smith, M. *The moral problem* (Blackwell, Oxford, 1994).

24. Zangwill, N. Aesthetic judgment. *The Stanford Encyclopedia of Philosophy (Fall 2014 Edition)*, ed Zalta, E. http://plato.stanford.edu/archives/fall2014/entries/aesthetic-judgment/ (2014).

25. Eres, R., Louis, W. & Molenberghs, P. Common and distinct neural networks involved in fMRI studies investigating morality: An ALE meta-analysis. *Soc Neurosci* **13,** 1-15 (2018).

26. Boccia, M. et al. Where does brain neural activation in aesthetic response to visual art occur? Meta/analytic evidence from neuroimaging studies. *Neurosci Biobehav R* **60,** 65-71 (2016).

27. Fontenelle, L. et al. Decoding moral emotions in obsessive-compulsive disorder. *NeuroImage* **19,** 82-9 (2018).

28. Wen, X. et al. The neural correlates of internal and external comparisons: an fMRI study. *Brain Struct Funct* **222,** 563-575 (2016).

29. Levy, D. & Glimcher, P. The root of all value: a neural common currency for choice. *Curr Opin Neurobiol* **22,** 1027-1038 (2012).

30. Sescousse, G., Li, Y. & Dreher, J. A common currency for the computation of motivational values in the human striatum. *Soc Cogn Affect Neur* **10,** 467-473 (2014).

31. Delgado, M. Reward-related responses in the human striatum. *Ann NY Acad Sci* **1104,** 70-88 (2007).

32. Peters, J. & Büchel, C. Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. *J Neurosci* **29,** 15727-34 (2009).

33. Howard, J., Gottfried, J., Tobler, P. & Kahnt, T. Identity-specific coding of future rewards in the human orbitofrontal cortex. *P Natl Acad Sci USA* **201503550** (2015).

34. Howard, J. & Kahnt, T. Identity-specific reward representations in orbitofrontal cortex are modulated by selective devaluation. *J Neurosci* **37,** 3473-16 (2017).

35. Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. Neural evidence for inequality-averse social preferences. *Nature*, **463**(7284), 1089-1091 (2010).

36. Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *P Natl Acad Sci*, **112**(25), 7851-7856 (2015).

37. Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E., & Dolan, R. J. Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, **85**(2), 418-428 (2015).

38. Maldjian, J., Laurienti, P., Kraft, R. & Burdette, J. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* **19,** 1233-1239 (2003).

39. Maldjian, J., Laurienti, P. & Burdette, J. Precentral gyrus discrepancy in electronic versions of the Talairach atlas. *Neuroimage* **21,** 450-455 (2004).

40. Hebart, M., Görgen, K., Haynes, J-D. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front Neuroinform* **8,** 1-17 (2015).

41. Chang, C. & Lin, C. LIBSVM: a library for support vector machines. *ACM T Intel Syst Tec* **2**, 23 (2011).

42. Fan, R., Chang, K., Hsieh, C., Wang, X. & Lin, C. LIBLINEAR: A library for large linear classification. *J Machine Learn Res* **9**, 1871-1874 (2008).

43. Oosterhof, N., Connolly, A., & Haxby, J. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics* (2016).

44. Stelzer, J., Chen, Y., & Turner, R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage* **65**, 69-82 (2013).

Figure Legends

Figure 1. Task and behavior. (a) The beginning of each trial specified judgment content and type. After 1.5s, a randomly selected image appeared below the specification. Participants had 4.5s to assess the image. They indicated and confirmed their judgment on a continuous rating scale (3s max). The orientation of the scale and the initial position of the cursor on the scale were randomized across trials, to prevent motor preparation during the image display phase. (b) Image ratings. The mean morality and aesthetics ratings for each of the 24 images were not correlated with one another. (c) Mean saccade number (top) and size (amplitude; bottom) during *image display* for each condition, averaged across images and participants. Error bars indicate standard deviations. There was no significant difference between the four judgment types for both saccade number and size.

Figure 2. Common univariate activations for morality and aesthetics judgments. Brain sections show regions activated for both the contrast (morality-speed) and the contrast (aesthetics-age), both whole-brain FWE corrected, $p<0.05$ and inclusively masked, thus representing above-threshold voxels for both contrasts. (a) Common cluster in frontal pole (extending into anterior cingulate cortex). (b) Contrast estimates at peak (0,50,13). (c) Common cluster in insula. (d) Contrast estimates at peak (-30,14,-17). (e) Common cluster in orbitofrontal cortex. (f) Contrast estimates at peak (-45,23,-11). All coordinates are in MNI space. Color bars indicate t-values, error bars standard error.

Figure 3. Specific univariate activation. Areas of activations for content-uncontrolled differences between morality and aesthetics judgments are shown in red, content-controlled differences in blue. Overlap is shown in pink. (a) Specificity for morality judgments. Regions activated for the contrast (morality-aesthetics) are shown in red and for (morality-speed)-(aesthetics-age) in blue. (b) Contrast estimates at overlap (-45,-61,19). (c) Specificity for aesthetics judgments. Regions activated for (aesthetics-morality) are shown in red and for (aesthetics-age)-(morality-speed) in blue. (d) Contrast estimates at overlap (24,-79,-5). All coordinates are in MNI space. Error bars indicate standard error of the mean.

Figure 4. Region of interest (ROI) analyses. (a) Morality ROI based on a meta-analysis (25). (b) Activations revealed by the contrast [(morality-speed)-(aesthetics-age)], p<0.05, FWE corrected. (c) Aesthetics ROI based on a meta-analysis (26). (d) Activations revealed by the contrast [(aesthetics-age)-(morality-speed)], p<0.05, FWE corrected. Abbreviations: am, anterior medial; dm, dorsomedial; vm, ventromedial; lTPJ/rTPJ, left/right temporoparietal junction; PFC, prefrontal cortex; post, posterior. Color bars represent t-values.

Figure 5. Common multivariate representations for morality and aesthetics judgments as revealed by cross-classification analyses. (a) In a first analysis, we trained a classifier to distinguish between morality and speed and tested it on aesthetics vs. age, *and vice versa*. All cross-classifications are significant at p<0.05, voxel-wise whole-brain FWE corrected (see Supplementary Table 6 for a full list). Medial and lateral prefrontal cortex (mPFC, LPFC), left temporoparietal

junction (TPJ), orbitofrontal cortex (OFC), and inferior temporal cortex (IT) show common representations of judgment types. (b) Test accuracies in peak voxels of the identified regions (chance level: 0.5). Error bars represent standard error, color bars t-values. (c) A second cross-classification analysis revealed common multivariate representation in medial PFC for morality and aesthetics judgments. Here we trained a classifier to distinguish between high and low morality ratings and tested it on high versus low aesthetics ratings, and *vice versa*. Cross-classification is significant at $p<0.05$, voxel-wise whole-brain FWE corrected. The figure shows the peak at MNI coordinates 6, 53, 25. This region overlapped with the medial prefrontal region identified in (a).

Figure 6. Specific multivariate representations for morality and aesthetics judgments. We trained a classifier to distinguish activity patterns for morality vs. aesthetics. The identified regions carried significant information about judgment type at $p<0.05$, voxel-wise whole-brain FWE corrected (see Supplementary Table 7 for a full list). (a) Activity patterns yielded by the morality vs. aesthetics classifier (red) identified specific multivariate representations in frontal and temporal regions. To assess whether these regions involved regions showing common multivariate representations, we replotted the data from the cross-classification analysis described in Figure 5a (blue). Overlap in temporoparietal junction (TPJ) and lateral prefrontal cortex (LPFC) shown in pink. (b) Test accuracies of the morality versus aesthetics classifier in peak voxels in TPJ and LPFC (chance level: 0.5). Error bars represent standard error.