MoralNet: Visual representations of moral intuitions in artificial and biological neural networks

Tim Lauer* (tl@leibniz-psychology.org)
Leibniz Institute for Psychology (ZPID), Trier, Germany

Karolina Drożdż* (karolina.drozdz@student.uva.nl)
University of Amsterdam, Amsterdam, The Netherlands

Sarah Müller (samu@leibniz-psychology.org)
Leibniz Institute for Psychology (ZPID), Trier, Germany

Frederic R. Hopp (fhopp@leibniz-psychology.org)
Leibniz Institute for Psychology (ZPID), Trier, Germany
Institute for Cognitive and Affective Neuroscience, Trier University, Trier, Germany
Trier University, Trier, Germany

*Shared first authorship

Abstract

Humans quickly detect moral situations in visual surroundings, suggesting that moral perception is attuned to features of the sensory environment. Yet, few computational models describe how combinations of stimulus features evoke perceptions of different moral situations. Here, we develop a convolutional neural network that decodes images into 10 distinct moral categories. We train and cross-validate the model on socio-moral images and show that image content is sufficient to predict the category of human morality ratings. In a fMRI study, we demonstrate that patterns of human visual cortex activity encode moral category–related model output and can decode multiple categories of moral scenes. Our preliminary results suggest that category-specific visual features can be reliably mapped to distinct moral intuitions, and they are coded in distributed representations within the human visual system.

Keywords: moral perception; social cognition; vision; mvpa; neuroAI; deep learning

Humans are vigilant moral observers, quickly and accurately detecting morally relevant situations in their environments (Gantman & Van Bavel, 2016). Moral Foundations Theory (MFT; Graham et al., 2013) postulates that moral perception is afforded by a set of evolved moral intuitions: rapid canonical responses to situations ancestrally linked to cooperation and group cohesion. If so, then moral intuitions may be directly elicited by features of the sensory environment. Indeed, morally relevant stimuli “pop-out” in early sensory processing (Decety & Cacioppo, 2012; Gantman et al., 2020), and multivariate pattern analysis (MVPA) in visual cortex can decode moral judgments of different moral actions (Hopp et al., 2023). Yet, few computational models describe how specific configurations of low-level (e.g., color) and high-level (e.g., objects) stimulus features evoke moral intuitions in sensory (e.g., visual) cortex. Here we draw on work by Kragel and colleagues (2019) who demonstrated that human ratings of emotions – a candidate source for moral intuitions – can be modeled solely from image features and decoded from neural activation patterns in the visual system. We tested three predictions: (H1) Models trained on image features alone can predict human ratings of moral intuitions; (H2) activation patterns in visual cortex are sufficient to decode moral intuitions; and (H3) visual representations in computational models of moral intuitions map onto distinct patterns of brain activity in visual cortex.

Results

Developing MoralNet (H1). Trained on 2,941 human-annotated socio-moral images (Crone et al., 2018), with 589 held out for testing, we developed “MoralNet,” a fine-tuned version of AlexNet (Krizhevsky, 2012) that classifies images into MFT’s ten moral foundations (five foundations split into virtue and vice categories) derived from annotators’ majority vote. Its performance was compared against four baseline models, each using a Histogram-Gradient Boosting Classifier trained on (i) high-level object features (1,000 object probabilities from a standard pretrained AlexNet), (ii) the number of people in the image (detected via YOLO), (iii) low-level color histograms, and (iv) low-level power spectral density (PSD). After cross-validation on the remaining 2,352 training images, MoralNet achieved an accuracy of .274 on the test set, surpassing all baseline models and slightly exceeding the object-features model (.236; see Figure 1a). The confusion matrix (Figure 1b) confirms correct predictions for every category, indicating that MoralNet’s performance was not dominated by a single class.

Decoding moral intuitions from fMRI data (H2). A representative subset of 120 images (Hopp et al., 2024) used to develop MoralNet was presented to 30 human participants in an event-related moral judgment fMRI experiment (3T multiband; TR 0.72; Figure 1c). Data were preprocessed using fMRIprep (Esteban et al., 2019) and participant-level beta maps for each image and moral category were created. We then used an MVPA approach with a linear SVM to perform pairwise decoding of the ten categories in the entire occipital lobe (Lancaster et al., 2000). In a leave-one-subject-out (LOSO) cross-validation analysis, all pairwise classification accuracies exceeded the chance level (Figure 1d).

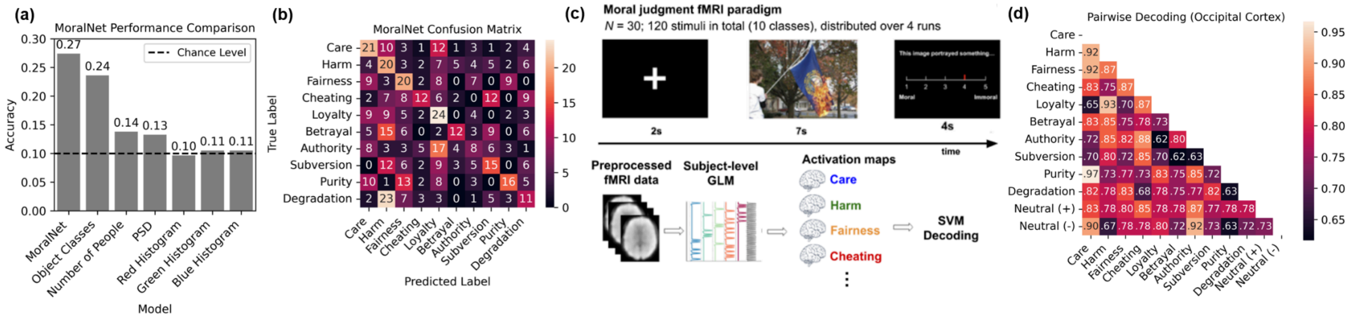

Figure 1: MoralNet's accuracy compared with six baseline models (a), confusion matrix on the test set (b), fMRI paradigm and preprocessing (c), and pairwise decoding in occipital cortex (d). Color bars represent the number of cases (b) and decoding accuracy (d).

Mapping computational representations to brain activity (H3). Finally, we examined whether the category-specific features learned by MoralNet are reflected in distributed patterns of human visual cortex activity. Using partial least squares (PLS) regression, we modeled activations in MoralNet's final fully connected layer based on fMRI responses in the occipital lobe to the same set of 120 images. We assessed the correlation between the predicted and observed activations across participants, using LOSO cross-validation. Across categories, a significant positive mean correlation was observed (r = .193 ± 0.009 SE, mean effect size d = 4.079, 80.51% of the noise ceiling, p < .001, permutation test). Results per category are shown in Table 1.

Discussion

First, our preliminary results confirm that models trained solely on image features can predict human moral intuitions, with MoralNet outperforming other baseline models. Notably, while low-level feature models showed minimal predictive utility, the high-level object-based model performed substantially better, suggesting that object identity may be important for moral classification. Second, multiple categories of moral intuitions could be reliably decoded from fMRI signals in the visual cortex, consistent with the notion that distributed patterns of activity in occipital areas encode both morally (Hopp et al., 2023) and emotionally (Kragel et al., 2019) relevant information. Finally, our demonstration that MoralNet's final-layer activations significantly align with visual cortex activity provides initial evidence that deep convolutional networks, trained exclusively on images, capture features relevant for neural representation of moral content—and that sensory cortical activity encodes these features.

Table 1. Category-level mapping from MoralNet to brain activity

Category	*r* ± SE	*d*	Noise Ceiling (%)
Care	.221 ± .017	2.34	78.21
Harm	.135 ± .019	1.309	73.51
Fairness	.244 ± .018	2.447	81.98
Cheating	.247 ± .02	2.276	78.55
Loyalty	.23 ± .015	2.735	81.96
Betrayal	.12 ± .017	1.259	76.5
Authority	.08 ± .02	0.725	58.07
Subversion	.191 ± .018	1.9	81.79
Purity	.238 ± .017	2.617	93.19
Degradation	.223 ± .017	2.375	87.89

Acknowledgments

F.R.H. acquired funding from the Amsterdam School of Communication Research (Grant No. ASCoR-u-2023-Hopp & ASCoR-u-2024-Hopp). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank Benjamin Jargow, Esmee Kouwen, and H. Steven Scholte for their assistance in data collection.

References

Crone, D. L., Bode, S., Murawski, C., & Laham, S. M. (2018). The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, *13*(1), e0190954.

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of neurophysiology*, *108*(11), 3068-3072.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, *16*(1), 111-116.

Gantman, A., Devraj-Kizuk, S., Mende-Siedlecki, P., Van Bavel, J. J., & Mathewson, K. E. (2020). The time course of moral perception: an ERP investigation of the moral pop-out effect. *Social Cognitive and Affective Neuroscience*, *15*(2), 235-246.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Hopp, F. R., Amir, O., Fisher, J. T., Grafton, S., Sinnott-Armstrong, W., & Weber, R. (2023). Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. *Nature Human Behaviour*, *7*(12), 2182-2198.

Hopp, F. R., Jargow, B., Kouwen, E., & Bakker, B. N. (2024). The Dutch moral foundations stimulus database: An adaptation and validation of moral vignettes and sociomoral images in a Dutch sample. *Judgment and Decision Making*, *19*, e10.

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science advances*, *5*(7), eaaw4358.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., ... & Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human brain mapping*, *10*(3), 120-131.