Visual moral inference and communication

Warren Zhu (warrenz@cs.toronto.edu), Aida Ramezani (armzn@cs.toronto.edu)
Department of Computer Science
University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program
University of Toronto

Abstract

Humans can make moral inferences from multiple sources of input. In contrast, automated moral inference in artificial intelligence typically relies on language models with textual input. However, morality is conveyed through modalities beyond language. We present a computational framework that supports moral inference from natural images, demonstrated in two related tasks: 1) inferring human moral judgment toward visual images and 2) analyzing patterns in moral content communicated via images from public news. We find that models based on text alone cannot capture the fine-grained human moral judgment toward visual stimuli, but language-vision fusion models offer better precision in visual moral inference. Furthermore, applications of our framework to news data reveal implicit biases in news categories and geopolitical discussions. Our work creates avenues for automating visual moral inference and discovering patterns of visual moral communication in public media.

Keywords: moral inference; multimodal fusion; language model; computer vision; artificial intelligence

Introduction

Morality plays a fundamental role in human cognition, but can machines tell right from wrong? This problem of computational moral inference, or automated inference of moral values, has been a topic of increasing relevance to artificial intelligence (AI) over the past decade (Abdulhai et al., 2023; Emelin et al., 2021; Forbes et al., 2020; Hämmerl et al., 2022; Hendrycks et al., 2021; Hoover et al., 2018; Jiang et al., 2021; Xie et al., 2020) pertaining to critical issues ranging from moral decision-making in autonomous vehicles (Awad et al., 2018) to alignment of AI and humans in cultural moral norms and values (Ramezani & Xu, 2023; Tao et al., 2024).

The dominant approach to moral inference relies on text and considers language as the sole medium for moral communication. For example, work from natural language processing and computational social science has developed text-based methods for automatic moral inference such as classifying moral sentiment (Garten et al., 2016; Johnson & Goldwasser, 2018, 2019; Lin et al., 2018; Pavan et al., 2020; Rezapour, Shah, & Diesner, 2019; Trager et al., 2022), detecting moral sentiment change (Garten et al., 2016; Mooijman et al., 2018; Ramezani et al., 2021; Xie et al., 2019), generating moral text generation and aligning language models with human values (Ammanabrolu et al., 2022; Emelin et al., 2021; Forbes et al., 2020; Hendrycks et al., 2020; Jiang et al., 2021; Shen et al., 2022).

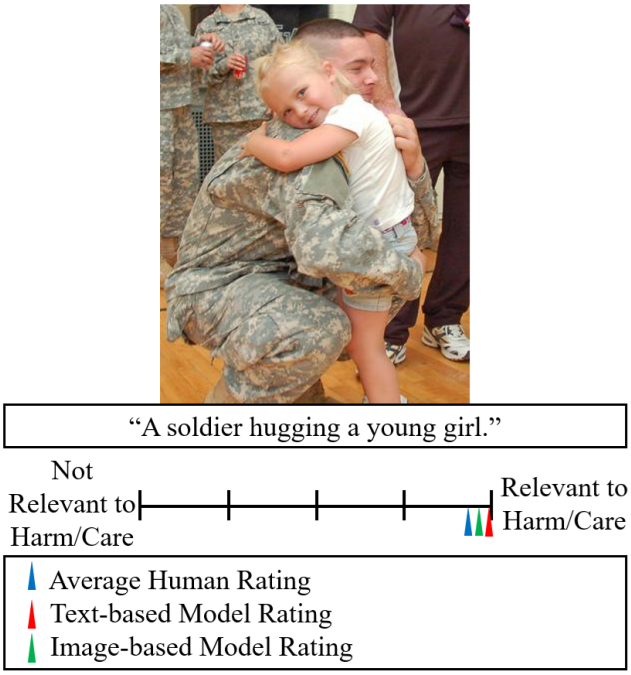

Figure 1: The photograph of a soldier hugging a child from Socio-Moral Image Database database used for analysis, accompanied by a caption generated by AZURE AI. The text-based and image-based representations of the image and its caption are both highly related to the moral foundation concerning Care, with the image-based representation offering a more accurate estimation of the moral content of the image based on the ground-truth human ratings.

This line of work has shown promise for textual moral inference. However, morality is often conveyed through mediums beyond language. In particular, prior research shows that images (e.g., photographs) can convey moral information beyond words and facilitate human ethical reasoning (Coleman, 2006). Figure 1 illustrates how a moral sense of care can be conveyed through our visual perception of a photograph. In this case, text-based moral inference methods may not be sufficient to capture the moral sentiment elicited by visually processing the image. As such, a quest for visual moral inference is warranted. While existing work has proposed meth-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

ods for filtering immoral or sensitive parts of an image (Jeong et al., 2022; Park et al., 2023; Wu et al., 2020), our study is distinct in that we focus on understanding how morality is communicated through the visual modality and how computational methods can automatically infer fine-grained human moral judgment in natural images at scale.

By leveraging moral information sourced from both language and vision, we contribute a fusion-based framework that extends text-based approaches to moral inference. We apply our framework to images from news media to understand how news images might communicate moral information to the public. We develop our framework to address the following two related problems:

- **Visual moral inference**: Can computational models drawn from AI make reliable prediction about fine-grained human moral judgment toward natural images such as photos?
- **Visual moral communication**: Can these models of visual moral inference be applied to analyzing how morals are embedded and communicated to the public through images, such as those appearing in news articles?

To address the first problem, we develop supervised models for visual moral inference using a large database of photographic images where human moral ratings are available.[1] This development allows us to tackle the second problem in an unsupervised way with minimal human intervention, which is useful because human moral ratings are typically scarce or unavailable for public images.

In developing our framework, we draw on the Socio-Moral Image Database (SMID) (Crone et al., 2018), the largest standardized visual moral stimulus set to our knowledge. This database provides a wide range of publicly-available photographic images, rated by humans for their moral content. The database includes ratings that capture the extent to which an image is related to the moral foundations based on the Moral Foundations Theory (Graham et al., 2011). Moral Foundations Theory is a modern theory of morality proposing that morality is modular and depends on five core foundations: *Care* (concerning the prevention/alleviation of suffering), *Fairness* (concerning the identification of exploitation/cheating), *Ingroup* (concerning self-sacrifice for the benefit of a collective), *Authority* (concerning respecting/obeying superiors), and *Purity* (concerning the avoidance of pathogens through the regulation of one's sexual and eating behaviors). For instance, as shown in Figure 1, the image of a soldier hugging a child is rated highly by human participants with respect to the Care moral foundation, while an image of a child respecting their parents would be rated highly for the Authority foundation. We demonstrate that combining visual image representations with textual representations (derived from image captions) tend to offer the most accurate prediction of fine-grained human moral ratings in this database.

Building on the visual moral inference framework, we show that it can be applied to exploring how moral information is communicated through images in a public news media. While prior work has investigated how morality and moral foundations, in particular, are communicated in various forms of text-based media, including news, social media, and child speech (Fulgoni et al., 2016; Hofmann et al., 2022; Hoover et al., 2020; Ramezani et al., 2021, 2022; Roy & Goldwasser, 2020; Roy et al., 2022; Shahid et al., 2020; Trager et al., 2022), how morals are communicated visually has not been comprehensively explored. Here we investigate visual moral communication using the New York Times images during the period 2010-2018 (Biten et al., 2019). We find evidence for implicit biases based on the moral content in the images displayed across different news categories and geographical regions. Here we summarize the landscape of the literature relevant to the development of our framework.

Computational approaches to moral inference. There has been a growing interest in using AI and computational methodologies for large-scale inference and discovery of moral values. The computational models developed under this paradigm of textual moral inference are typically grounded in established theories of human morality, and they have been applied to classify text based on different moral categories, particularly moral foundations (Asprino et al., 2022; Garten et al., 2016; Hoover et al., 2020; Johnson & Goldwasser, 2018, 2019; Kobbe et al., 2020; Lin et al., 2018; Liscio et al., 2022; Mooijman et al., 2018; Pavan et al., 2020; Qian et al., 2021; Rezapour, Ferronato, & Diesner, 2019; Rezapour, Shah, & Diesner, 2019; Roy & Goldwasser, 2021; Santos & Paraboni, 2019; Trager et al., 2022). One important takeaway from this line of research suggests that people often communicate their moral views through language, and language use or language modelling can be an effective approach for making inferences about human morals at scale.

With recent advance in generative language models, a new research area has emerged that evaluates and aligns the outputs of these models to human (moral) values. Efforts in this area vary from assessing and fine-tuning language-model generated text according to different moral theories (e.g., Virtue ethics, Utilitarianism) (Abdulhai et al., 2023; Hendrycks et al., 2020; Jin et al., 2024; Simmons, 2022) to grounding model outputs in crowd-sourced generated datasets of moral norms (Ammanabrolu et al., 2022; Bai et al., 2022; Emelin et al., 2021; Forbes et al., 2020; Jiang et al., 2021; Liu et al., 2022; Shen et al., 2022).

Our study extends the existing research on moral inference from text to the visual domain and takes an initial step toward developing fusion models for moral inference.

Psychological and cognitive studies on visual moral communication. Images can contain information that can go beyond and even be different from their textual description (Coleman, 2006; Lang et al., 2015). Research on visual

[1]The code to replicate our framework and analysis can be found in the following repository: https://github.com/CoderWarren/Visual-Moral-Inference-and-Communication/tree/main.

moral communication suggests certain symbolic depictions of icons and events, such as cigarettes (Yang et al., 2018), can significantly influence people's perception of morality (De Freitas & Alvarez, 2018). For example, colors of black and white symbolize purity and pollution for a lot of people (Frank & Gilovich, 1988; Meier et al., 2004; Sherman & Clore, 2009; Stabler & Johnson, 1972), and certain choices of these colors can even affect people's moral judgments in other modalities such as text (Zarkadi & Schnall, 2013).

Our work builds on these existing empirical studies toward automated analysis of visual moral communication, and we demonstrate how our framework can be used for detecting implicit patterns in public communication of moral information from news images.

Language-vision fusion models. Besides a multitude of language models such as BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019)) and vision models (such as YOLO (Redmon et al., 2016) and SAM (Kirillov et al., 2023), there are also various vision-language models (such as FLAVA (Singh et al., 2022) and BLIP (Li et al., 2022)), some of which have been used to automate image captioning (Hu et al., 2023; Li et al., 2023). CLIP (Radford et al., 2021), in particular, is a powerful vision-language model trained on 400 million image-text pairs that is capable of generating robust image and textual embeddings for downstream tasks. Although CLIP text encoders under-perform SBERT encoders (Reimers & Gurevych, 2019a) on general natural language understanding tasks, CLIP textual encoders possess a unique ability that SBERT encoders do not—they can associate a text and its visual appearance, which is more similar to human perception (Chen et al., 2023). Additionally, CLIP has been used in many in state-of-the-art Vision-Language models such as DALL-E (Ramesh et al., 2021) and SAM (Kirillov et al., 2023).

For our work, we consider CLIP for visual moral inference and systematically compare it to different approaches for building a reliable computational framework that integrates images and text for reliable and automated moral inference.

Data

Socio-Moral Image Database (SMID)

We use the Socio-Moral Image Database (Crone et al., 2018). This dataset consists of 2,941 photographic images with their normative ratings (each on a 1-5 scale) for morality ("blameworthy" to "praiseworthy"), and relevance to the five moral foundations ("unrelated" to "related") from around 2,000 human participants. By itself, SMID does not feature captions for the images, we therefore generated captions for the aforementioned images through the use of Microsoft's AZURE AI (Version 4.0, accessed through their 2024-02-01 REST API). [2] To ensure robustness of our results, we also used GOOGLE VERTEX AI as an alternative to AZURE AI. Given the guardrails of VERTEX AI, we were unable to generate captions for some sensitive images, so we therefore used AZURE AI for all the analyses reported here.

[2]The model can be found at https://azure.microsoft.com/en-us/solutions/ai

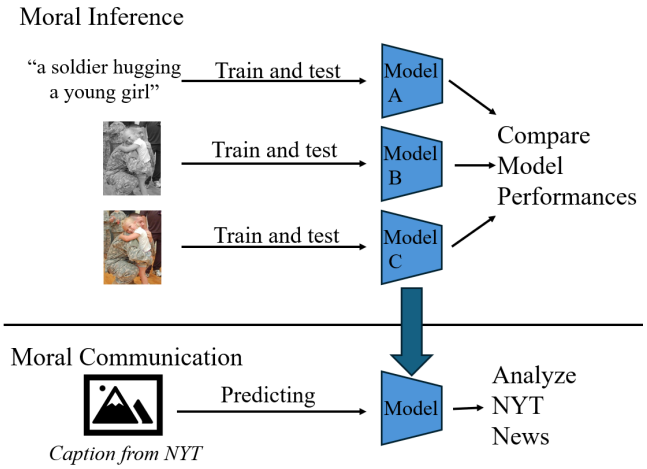

Figure 2: An illustration of our image-text fusion framework for visual moral inference and communication. Top plot: Evaluation of different text and image representations of the input figures used to train computational models for moral inference. Bottom plot: Applying the text-image fusion model to uncover implicit patterns of visual moral communication in news media.

GoodNews New York Times images

GoodNews (Biten et al., 2019) is a dataset of more than 466,000 New York Times images from January 2010 to June 2018, collected for image captioning. Along with images, GoodNews also provides image captions and links to their respective articles. Unlike the SMID data, images of the NYT news data had no human moral ratings. The framework we built allows one to estimate moral ratings for this news source but also flexibly for other sources of data where human ratings are not available or accessible.

Computational methodology

In this section, we provide details for the construction of our framework that comprises of visual moral inference and communication, summarized and illustrated in Figure 2.

Models of visual moral inference

We compare different representations of images in SMID to evaluate their ability to capture the moral content in images. Our analysis incorporates both textual (from image captions) and visual information, and we consider the text-based models as baselines. We summarize the different model classes as follows:

- **Bag-of-Words (BoW):** We use the Bag-of-Words representation of image captions as our baseline for moral inference. Stopwords identified by `NLTK` were removed. This approach captures how individual words, irrespective

of their position and context, reflect human moral concerns. Although simple and less accurate than the other approaches, this baseline offers interpretability at the token level and insights into the topics that trigger the most morally relevant responses in human participants.

- **Contextual Embedding:** We use the `ViT-B/32` CLIP text-encoder (Radford et al., 2021) and a variety of SBERT encoders (Reimers & Gurevych, 2019b) to generate contextual representation of image captions. The full list of SBERT encoders can be found in Table 1. The CLIP text-encoder model allows us to capture how image captions are associated with visual appearances, while SBERT encoders provide better quality language representation (Chen et al., 2023).

- **Grayscaled Image Embedding:** We produce the corresponding grayscaled representation of the images with the purpose of studying the role of color in visual moral inference. The images are then transformed into high-dimensional representations using the `ViT-B/32` CLIP image-encoder.

- **Original Image Embedding:** These embeddings were produced by running the original image through `ViT-B/32` CLIP image-encoder.

- **Joint Image-Text Embedding:** These embeddings were produced by adding the Original Image Embeddings to the CLIP Textual Embedding.

Our BoW vectors consisted of 1,579 dimensions, and CLIP embeddings were 512-dimensional. The size of our SBERT embeddings varied, with `all-MiniLM-L6-v` and `all-MiniLM-L12-v2` being 384-dimensional and the rest being 768-dimensional. Before further analysis, BoW vectors were normalized by dividing by the sum of its components, while embeddings were normalized by dividing by their Euclidean Norm. We divided the images into an 80(training)-20(prediction) split, and used their representations to train ridge regression models for each of the dimensions in SMID related to morality (i.e., Morality, and relevance to the five moral foundations). Ridge regression involves minimizing the objective loss function: $||y - XW||_2^2 + \alpha * ||W||_2^2$. Here, y represents image scores we wish to predict (e.g., Morality), X is the image representations (e.g., CLIP representations of the original images), and W is the regression parameter. We find the best α (penalty parameter) using grid search on the validation set. Specifically, we run 3 rounds of 10-fold cross-validation on the training set and choose the value that yields the best R^2 score on average. We chose a linear regression model over more complex approaches to compare how well the existing features (captured by the different embeddings) predict the target moral variables without further manipulation. After identifying the hyper-parameters and the most reliable representation, we trained our final model on full set of the SMID images.

Analysis of visual moral communication

For analyzing visual moral communication in public news images, we apply the best performing model drawn from the set of models described in the previous section on the New York Times article images as provided in the GoodNews database to obtain their estimated moral ratings. Unlike SMID images where we constructed the captions using automatic AI tools, here we use captions from the actual NYT articles.

GoodNews further provides the publication date and the news category the articles fall into. From these categories we extracted images of different regions (nyregion, us, world/europe, world/asia, world/africa, and world/middleeast) and article topics (business, health, sports, technology, and science) for exploring differences and potential biases in moral communication under these categories.

Results

Evaluating models of visual moral inference

The overall performance of the models after hyperparameter optimization are summarized in Table 1. The joint image-text embedding produces the most reliable representation for predicting moral ratings with the average $R^2 = 0.6320$. This is closely followed by the original image representation with the average $R^2 = 0.6270$. Morality and the relevance to the Authority, Care, and Ingroup moral foundations were predicted more accurately than Fairness and Purity moral foundations by all models. Moreover, the grayscaled image representations performed slightly worse across all 6 variables of interest compared to our coloured image models, suggesting that similar to humans (Zarkadi & Schnall, 2013), colour also acts as a visual moral stimuli in CLIP representations. In all cases, the image-based representation produced by CLIP outperformed the text-based models, with the worst-performing visual model (grayscaled image representations) having a correlation of at least 0.10 higher for each of the 6 variables of interest compared to the best performing text-based model (CLIP-embedding textual model).

Among the text-based models, CLIP text representation slightly outperforms each of the SBERT model. While SBERT is shown to be better at natural language understanding tasks, our goal primarily revolves around replicating human ratings for images, which leverages CLIP's effectiveness in associating a text with its visual appearance (Chen et al., 2023). Despite the comparatively poor performance of the BoW model, we found that it was capable of detecting concepts and actions related to moral foundations, such as *soldier*, *police* and *saluting* for Authority and *crashed*, *destroyed*, and *explosion* for Care.

Our results show that the visual stimuli in these photographic images elicit moral responses from human participants that cannot be accurately predicted by relying solely on textual representations of images. Consequently, models that rely only on captions cannot reliably infer the actual moral ratings elicited by such images in participants. In sum-

Model	M	A(R)	F(R)	C(R)	I(R)	P(R)	Average
BoW	0.3747	0.4441	0.3049	0.3871	0.4146	0.3368	0.3770
SBERT (all-mpnet-base-v2)	0.4520	0.5011	0.3475	0.4550	0.4580	0.3628	0.4294
SBERT (multi-qa-mpnet-base-dot-v1)	0.4244	0.4932	0.3695	0.4483	0.4496	0.3417	0.4211
SBERT (all-distilroberta-v1)	0.4374	0.4973	0.3469	0.4588	0.4552	0.3371	0.4221
SBERT (all-MiniLM-L12-v2)	0.4276	0.4862	0.3435	0.4489	0.4700	0.3397	0.4193
SBERT (multi-qa-distilbert-cos-v1)	0.4281	0.5083	0.3529	0.4351	0.4544	0.3661	0.4242
SBERT (all-MiniLM-L6-v2)	0.4113	0.4806	0.3385	0.4253	0.4481	0.3405	0.4074
CLIP (Caption)	0.4306	0.5172	0.3519	0.4677	0.4696	0.3627	0.4333
CLIP (Grayscaled Image)	0.6245	0.6214	0.5416	0.5875	0.6095	0.5338	0.5864
CLIP (Colour Image)	0.6742	0.6560	**0.5796**	0.6360	0.6408	0.5757	0.6270
CLIP (Joint)	**0.6755**	**0.6701**	0.5704	**0.6513**	**0.6427**	**0.5822**	**0.6320**

Table 1: The R^2 scores between predicted and actual ratings of the SMID test set for the best performing models. Purely text-based models are found on the top section of the table–models involving images in some way are found on the bottom section. Here, M, A(R), F(R), C(R), I(R), and P(R) respectively stand for Morality, and relevance to Authority, Fairness, Care, Ingroup and Purity. The best R^2 scores for each predicted variable are shown in **bold**. The best scores using the text-based models are underlined.

mary, given that the CLIP model itself is developed with joint image-text training, our results demonstrate the effectiveness of incorporating visual stimuli with textual information for both moral and non-moral (i.e., valence and arousal) prediction based on fine-grained human judgment.

Visual moral communication in public news

To investigate how moral values may be embedded and communicated through visual stimuli in public media, we use our best performing model (i.e., joint image-text model) on the article images in the New York Times articles published in 2010-2018, using the GoodNews dataset (Biten et al., 2019).

Public news images contain implicit moral biases across regions. We find evidence for moral bias from images included in NYT news categorized across different regions: nyregion, us, world/europe, world/asia, world/africa and world/middleeast. As shown in Figure 3, images in nyregion (region of New York) and us (USA) exhibit more morally positive (i.e., moral) information and sentiment compared to the images in world/europe, world/asia, world/africa, and world/middleeast. Moreover, images in the world/africa and world/middleeast regions are associated more strongly with Care and Purity moral foundations compared to the other regions, as most NYT images associated with world/africa and world/middleeast typically depict warfare and the aftermath of combat. Examples of captions of these images include "Children receiving treatment after the gas attack" and "Rebels with the body of the commander".

Different news categories engage different moral foundations. Table 2 displays the average predicted morality for news categories health, sports, business, science, and technology. Health has the highest average Morality score with many images of people and animals being treated in clinical settings.

Furthermore, Figure 4 displays the average relevance to each of the different moral foundations moral scores. We find that the five categories exhibit different moral foundations averaged across all years. For example, sports, with many images of referees, engages the Fairness moral foundation more so than other categories. This category is also associated with the Ingroup moral foundation, with images showing sports fans, teammates, and families banding together. Among these categories, images in health appeal the most to the Care moral foundation. This is followed by science, which also includes images depicting clinical treatment. Health also consists of many images of people praying, possibly making it most relevant to the Purity moral foundation across all five categories.

Category	Morality
health	**3.2821 ($\pm$ 0.0069)**
sports	3.1177 ($\pm$ 0.0015)
business	3.0967 ($\pm$ 0.0020)
science	3.0884 ($\pm$ 0.0050)
technology	3.0880 ($\pm$ 0.0038)

Table 2: The mean predicted Morality scores for each of our categories, along with the standard errors of the mean.

Bootstrap testing was conducted to verify that these results are not based on random chance. We found strong evidence against health having the highest average Morality score ($P < 0.01$) and relevance to the Care ($P < 0.0001$) and Purity ($P < 0.001$) simply due to random chance. Evidence was also found against sports having the highest relevance to Fairness ($P < 0.0001$) simply due to chance.

Discussion and conclusion

Automated moral inference is an area of increasing relevance to AI and it has a foreseeable impact on empirical research in morality. We show that moral inference solely based on text

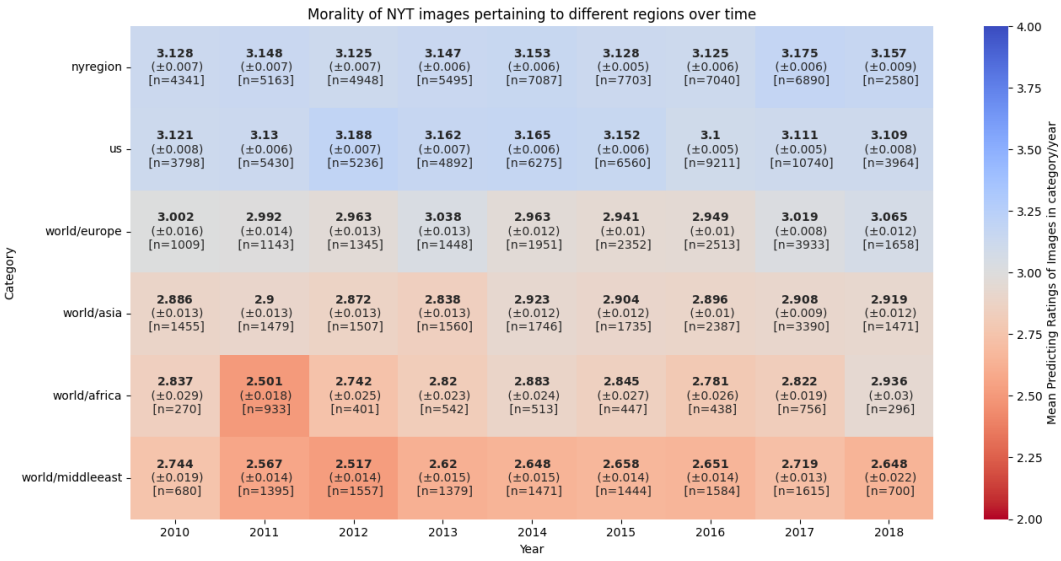

Figure 3: The predicted Morality scores of images corresponding to each regional category. Within each cell, the mean Morality score has been **bolded** on top, the standard errors of each mean are within the parentheses in the middle, and the number of images matching each year/category are in the square brackets on the bottom.

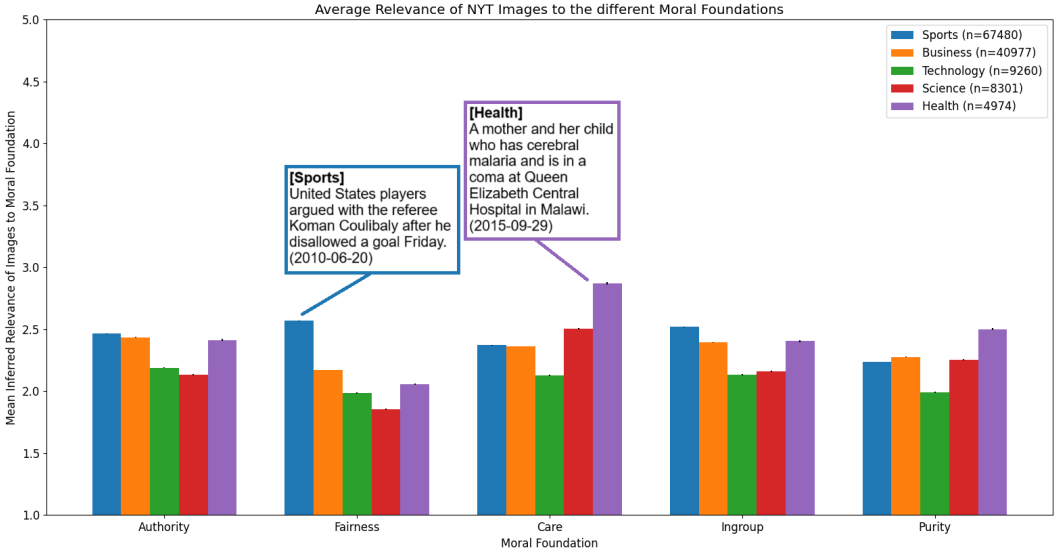

Figure 4: The mean predicted relevance to different moral foundations across all years for the news categories of interest. A rating of 1 indicates that an image is unrelated to the moral foundation, while a rating of 5 indicates that an image is highly related to the moral foundation. The number of images in each category can be found in the top right corner. Error bars indicate the standard error of the mean. Captions of sample images with high moral relevance are shown above the corresponding bars.

is inadequate for understanding the diverse mediums through which morals may be communicated. We found evidence that text can strengthen moral inference from visual stimuli, but text derived from images alone is not sufficient to capture the fine-grained moral sentiment toward images. Previous research has shown that visual perception can influence one's moral interpretation of text (Coleman, 2006; Zarkadi & Schnall, 2013), and we have fulfilled a need to developing methods for reconstructing fine-grained human moral judgment of natural images. We also demonstrated the utility of our framework to identifying implicit biases in images taken from public news. Future work may extend our framework to accommodate other sources of information such as sound and audios, movies and videos, and therefore exploring how different modalities may reflect or affect moral judgment. We hope that our work will generate further research in integrating language and vision with other modalities toward a more holistic and multimodal approach to moral inference.

Acknowledgments

This research is supported by an Ontario Early Researcher Award #ER19-15-050 to YX.

References

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2023). Moral foundations of large language models. *The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R2HCAI)*.

Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., & Choi, Y. (2022). Aligning to Social Norms and Values in Interactive Narratives. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5994–6017.

Asprino, L., Bulla, L., De Giorgis, S., Gangemi, A., Marinucci, L., & Mongiovi, M. (2022). Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 33–41.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *ArXiv preprint arXiv:2212.08073*.

Biten, A. F., Gomez, L., Rusiñol, M., & Karatzas, D. (2019). Good news, everyone! context driven entity-aware captioning for news images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12458–12467.

Chen, Z., Chen, G., Diao, S., Wan, X., & Wang, B. (2023, July). On the difference of BERT-style and CLIP-style text encoders. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 13710–13721). Association for Computational Linguistics.

Coleman, R. (2006). The effects of visuals on ethical reasoning: What's a photograph worth to journalists making moral decisions? *Journalism & Mass Communication Quarterly*, *83*(4), 835–850.

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), 1–34.

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, *178*, 133–146.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., & Choi, Y. (2021). Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 698–718.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020). Social chemistry 101: Learning to reason about social and moral norms. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670.

Frank, M., & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of personality and social psychology*, *54*, 74–85.

Fulgoni, D., Carpenter, J., Ungar, L., & Preoţiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3730–3736.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366.

Hämmerl, K., Deiseroth, B., Schramowski, P., Libovickỳ, J., Fraser, A., & Kersting, K. (2022). Do Multilingual Language Models Capture Differing Moral Norms? *ArXiv preprint arXiv:2203.09904*.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning AI with shared human values. *9th International Conference on Learning Representations (ICLR)*.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. *International Conference on Learning Representations*.

Hofmann, V., Dong, X., Pierrehumbert, J., & Schuetze, H. (2022, July). Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Findings of the association for computational linguistics: Naacl 2022* (pp. 536–550). Association for Computational Linguistics.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., Dehghani, M., & Donnellan, M. B. (2018). Moral framing and charitable donation: Integrating exploratory social media analyses

and confirmatory experimentation. *Collabra: Psychology*, *4*(1).

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral Foundations Twitter Corpus: A collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hu, J., Cavicchioli, R., & Capotondi, A. (2023). Exploiting multiple sequence lengths in fast end to end training for image captioning. *2023 IEEE International Conference on Big Data (BigData)*, 2173–2182.

Jeong, Y., Park, S., Moon, S., & Kim, J. (2022). Zero-shot visual commonsense immorality prediction. *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., & Choi, Y. (2021). Delphi: Towards Machine Ethics and Norms. *ArXiv*, *abs/2110.07574*.

Jin, Z., Levine, S., Gonzalez, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., & Schölkopf, B. (2024). When to make exceptions: Exploring language models as accounts of human moral judgment. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Johnson, K., & Goldwasser, D. (2018). Classification of moral foundations in microblog political discourse. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 720–730.

Johnson, K., & Goldwasser, D. (2019, June). Modeling behavioral aspects of social media discourse for moral classification. In S. Volkova, D. Jurgens, D. Hovy, D. Bamman, & O. Tsur (Eds.), *Proceedings of the third workshop on natural language processing and computational social science* (pp. 100–109). Association for Computational Linguistics.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. B. (2023). Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003.

Kobbe, J., Rehbein, I., Hulpuș, I., & Stuckenschmidt, H. (2020, December). Exploring morality in argumentation. In E. Cabrio & S. Villata (Eds.), *Proceedings of the 7th workshop on argument mining* (pp. 30–40). Association for Computational Linguistics.

Lang, A., Bailey, R., & Connolly, S. R. (2015). Encoding systems and evolved message processing: Pictures enable action, words enable thinking. *Media and Communication*, *3*(1), 34–43.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*.

Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*.

Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., & Ji, H. (2018). Acquiring background knowledge to improve moral value prediction. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 552–559.

Liscio, E., Dondera, A., Geadau, A., Jonker, C., & Murukannaiah, P. (2022, July). Cross-domain classification of moral values. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Findings of the association for computational linguistics: Naacl 2022* (pp. 2727–2745). Association for Computational Linguistics.

Liu, R., Zhang, G., Feng, X., & Vosoughi, S. (2022). Aligning Generative Language Models with Human Values. *Findings of the Association for Computational Linguistics: NAACL 2022*, 241–252.

Meier, B. P., Robinson, M. D., & Clore, G. L. (2004). Why good guys wear white: Automatic inferences about stimulus valence based on brightness [PMID: 14738513]. *Psychological Science*, *15*(2), 82–87.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396.

Park, S., Moon, S., & Kim, J. (2023). Ensuring visual commonsense morality for text-to-image generation [arXiv:2212.03507].

Pavan, M. C., dos Santos, W. R., & Paraboni, I. (2020). Twitter moral stance classification using long short-term memory networks. *Brazilian Conference on Intelligent Systems*, 636–647.

Qian, M., Laguardia, J., & Qian, D. (2021). Morality beyond the lines: Detecting moral sentiment using ai-generated synthetic context. *International Conference on Human-Computer Interaction*, 84–94.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation.

Ramezani, A., Liu, E., Ferreira Pinto Junior, R., Lee, S. W., & Xu, Y. (2022). The emergence of moral foundations in child language development. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Ramezani, A., & Xu, Y. (2023). Knowledge of cultural moral norms in large language models. *Proceedings of the 61th*

Annual Meeting the Association for Computational Linguistics: ACL.

Ramezani, A., Zhu, Z., Rudzicz, F., & Xu, Y. (2021). An unsupervised framework for tracing textual sources of moral change. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1215–1228.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Reimers, N., & Gurevych, I. (2019a, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019b). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Rezapour, R., Ferronato, P., & Diesner, J. (2019). How do moral values differ in tweets on social movements? *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 347–351.

Rezapour, R., Shah, S. H., & Diesner, J. (2019, June). Enhancing the measurement of social effects by capturing morality. In A. Balahur, R. Klinger, V. Hoste, C. Strapparava, & O. De Clercq (Eds.), *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 35–45). Association for Computational Linguistics.

Roy, S., & Goldwasser, D. (2020, November). Weakly supervised learning of nuanced frames for analyzing polarization in news media. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7698–7716). Association for Computational Linguistics.

Roy, S., & Goldwasser, D. (2021, June). Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the ninth international workshop on natural language processing for social media* (pp. 1–13). Association for Computational Linguistics.

Roy, S., Nakshatri, N. S., & Goldwasser, D. (2022). Towards few-shot identification of morality frames using in-context learning. *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, 183–196.

Santos, W., & Paraboni, I. (2019). Moral stance recognition and polarity classification from twitter and elicited text. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1069–1075.

Shahid, U., Di Eugenio, B., Rojecki, A., & Zheleva, E. (2020, July). Detecting and understanding moral biases in news. In C. Bonial, T. Caselli, S. Chaturvedi, E. Clark, R. Huang, M. Iyyer, A. Jaimes, H. Ji, L. J. Martin, B. Miller, T. Mitamura, N. Peng, & J. Tetreault (Eds.), *Proceedings of the first joint workshop on narrative understanding, storylines, and events* (pp. 120–125). Association for Computational Linguistics.

Shen, T., Geng, X., & Jiang, D. (2022, October). Social norms-grounded machine ethics in complex narrative situation. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 1333–1343). International Committee on Computational Linguistics.

Sherman, G. D., & Clore, G. L. (2009). The color of sin: White and black are perceptual symbols of moral purity and pollution. *Psychological Science*, *20*(8), 1019–1025.

Simmons, G. (2022). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *ArXiv preprint arXiv:2209.12106*.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15617–15629.

Stabler, J. R., & Johnson, E. E. (1972). The meaning of black and white to children. *International Journal of Symbology*, *3*(3), 11–21.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS nexus*, *3*(9), pgae346.

Trager, J., Ziabari, A. S., Davani, A. M., Golazazian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., Reimer, N. K., Reyes, M., et al. (2022). The Moral Foundations Reddit Corpus. *ArXiv preprint arXiv:2208.05545*.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020, September). Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision* (pp. 322–339). Springer.

Xie, J. Y., Ferreira Pinto Junior, R., Hirst, G., & Xu, Y. (2019, November). Text-based inference of moral sentiment change. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4654–4663). Association for Computational Linguistics.

Xie, J. Y., Hirst, G., & Xu, Y. (2020). Contextualized moral inference. *ArXiv preprint arXiv:2008.10762*.

Yang, S., Maloney, E. K., Tan, A. S. L., & Cappella, J. N. (2018). When Visual Cues Activate Moral Foundations: Unintended Effects of Visual Portrayals of Vaping within Electronic Cigarette Video Advertisements. *Human Communication Research*, *44*(3), 223–246.

Zarkadi, T., & Schnall, S. (2013). "Black and white" thinking: Visual contrast polarizes moral judgment. *Journal of Experimental Social Psychology*, *49*(3), 355–359.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019, July). ERNIE: Enhanced language representation with informative entities. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1441–1451). Association for Computational Linguistics.