Building Interpretable Models for Moral Decision-Making

Mayank Goel[1,*], Aritra Das[2] and Paras Chopra[1]

[1]Lossfunk
[2]Ashoka University

Abstract

We build a custom transformer model to study how neural networks make moral decisions on trolley-style dilemmas. The model processes structured scenarios using embeddings that encode who is affected, how many people, and which outcome they belong to. Our 2-layer architecture achieves 77% accuracy on Moral Machine data while remaining small enough for detailed analysis. We use different interpretability techniques to uncover how moral reasoning distributes across the network, demonstrating that biases localize to distinct computational stages among other findings.

Keywords

Mechanistic Interpretability, Machine Ethics, Transformers

1. Introduction

As AI systems increasingly confront moral dilemmas, from autonomous vehicles to content moderation, we face an urgent problem to understand not just what these systems decide but also how they reason and where moral biases originate within their architectures. We train a small model (104K Parameters) which can learn general moral principles and can be trained on 3M scenarios from the Moral Machine Dataset [1]. To address this challenge, we establish that moral principles must be encoded in the model architecture itself. Rather than probing opaque large language models post-hoc, we build a custom architecture from the ground up. Each token in our architecture represents three pieces of information: the character type, the number present in the scenario, and the team assignment. This representation encodes our hypothesis that moral computation reduces to identifying stakeholders, quantifying impact, and resolving competing interests. Further, we employ complementary interpretability methods to understand the model's moral reasoning process. We use causal intervention to quantify the influence of individual character types on decisions. Additionally, we apply layer-wise attribution to trace how different moral biases localize to specific computational stages across the network. Finally, we use circuit probing to identify sparse subnetworks that causally implement the final moral scores. Overall, our work makes the following main contributions:

- Firstly, we design a 2-layer transformer specifically optimized for moral reasoning tasks, achieving 77% accuracy on Moral Machine trolley problems. Our architecture demonstrates that transparency and predictive capability in moral choices are not mutually exclusive.
- Secondly, we establish an exploratory interpretability analysis of moral decision-making combining causal intervention, layer-wise attribution, and circuit probing to map the internal mechanisms underlying moral judgment in transformers.

2. Related Work

The Moral Machine dataset [1] was originally a crowdsourced collection of 40 million human decisions on AV dilemmas. Recently, it has been used to evaluate the ethical reasoning capabilities of large language models (LLMs). Jin et al. [2] translated 1,000 trolley-problem vignettes into over 100 languages

AAAI 2026 Workshop on Machine Ethics: from formal methods to emergent machine ethics, January 26th, 2026, Singapore

[*]Corresponding author.

mayank.goel@lossfunk.com (M. Goel); aritra.das@ashoka.edu (A. Das); paras@lossfunk.com (P. Chopra)

and compared LLM responses to cross-cultural human norms. bin Ahmad and Takemoto [3] evaluated 52 LLMs (GPT, Claude, Llama, etc.) on Moral Machine scenarios, finding larger models aligned more closely with human aggregate preferences. In parallel, several datasets have been developed to model moral and social norms. The ETHICS dataset [4] spans justice, virtue, duty, and commonsense, shows the partial success of Language Models at value-aligned judgment. Social Chemistry 101 [5] provides 100K situations annotated with normative "rules of thumb," and has been used to train models to generalize commonsense moral reasoning. Resources based on Moral Foundations Theory [6] enable categorization of ethical text across key moral dimensions (e.g., care, fairness, loyalty). Mechanistic interpretability studies have begun to explore how transformer-based models encode moral reasoning, with work identifying distinct "moral neurons" activating on ethical stimuli [7] and applying PCA and probing techniques to isolate moral subspaces in representation space [8].

3. Architecture

We propose a transformer-based model designed specifically for moral reasoning over structured scenario representations. Unlike prior work that applies general-purpose LLMs to ethical dilemmas, our architecture exploits the compositional structure inherent in trolley-problem scenarios.

3.1. Input Representation

Each moral dilemma is represented as a pair of outcomes, where an outcome is a vector over 23 character types (e.g., Man, Woman, Criminal, Doctor). Given a scenario with outcomes $\mathcal{O}_0$ and $\mathcal{O}_1$, we encode each outcome as a sequence of tokens—one per character type.

For every character c with cardinality n_c in outcome $\mathcal{O}_t$ (where $t \in \{0, 1\}$ denotes the team, one or the other), we construct a compositional embedding:

$$\mathbf{e}_c^{(t)} = [\mathbf{E}_{\text{char}}(c) \, ; \, \mathbf{E}_{\text{card}}(n_c) \, ; \, \mathbf{E}_{\text{team}}(t)] \tag{1}$$

where $[\cdot \, ; \, \cdot]$ denotes concatenation, and the embedding dimensions are allocated as $d_{\text{char}} = d/2$, $d_{\text{card}} = d_{\text{team}} = d/4$, summing to total embedding dimension d.

This design choice reflects our hypothesis that moral reasoning decomposes into: (1) *who* is affected (character identity), (2) *how many* are affected (cardinality), and (3) *which side* they belong to (team membership).

3.2. Transformer Processing

We prepend a learnable [CLS] token to the sequence of 46 character tokens (23 per outcome) and process the full sequence with a standard transformer encoder using pre-norm residual connections. The architecture uses $L = 2$ layers and $H = 2$ attention heads with embedding dimension $d = 64$. The relatively shallow architecture is motivated by the structured nature of our input: unlike natural language, our scenarios do not require deep hierarchical composition.

The transformer learns to compute cross-outcome comparisons via self-attention. Critically, the team embeddings allow the model to form separate representations of the two outcomes before computing their relative moral value through attention interactions.

3.3. Classification Head

The final [CLS] token representation is passed through a two-layer MLP with GELU activation to produce a scalar logit. This logit represents the model's preference for outcome 1 over outcome 0, which we convert to probabilities via sigmoid. The design mirrors recent findings in mechanistic interpretability [7] that moral reasoning in transformers localizes to specific aggregation points in the representation space. This probabilistic design naturally handles conflicts in the training set: contradictory preferences push the model toward intermediate probabilities, allowing it to learn which scenario features produce genuine moral uncertainty rather than being forced to memorize arbitrary choices.

3.4. Symmetry and Invariance

A key property of moral judgments is *side-invariance*: the choice between saving group A versus group B should be the complement of the choice between saving group B versus group A. To enforce this, we apply a symmetrization procedure at inference:

$$p(\mathcal{O}_1 \succ \mathcal{O}_0) = \frac{1}{2}\left[\sigma(f(\mathcal{O}_0, \mathcal{O}_1)) + \left(1 - \sigma(f(\mathcal{O}_1, \mathcal{O}_0))\right)\right] \tag{2}$$

where f is our model, σ is the sigmoid function, and $\mathcal{O}_1 \succ \mathcal{O}_0$ denotes preferring outcome 1. This averages predictions from both orderings, guaranteeing consistent probability assignments regardless of input order. Another motivation for this is the lack of symmetry we see at inference time, motivating an invariant inference for better predictability.

3.5. Model Configuration

We explored several architecture configurations to balance representational capacity with model compactness. We use a subset of the Moral Machine data which had only people who filled out the survey form, which led to a dataset of 5.4M, of which 1.7M were unique scenarios and kept aside as the validation set, and the remaining were used as the training set. By using only unique situations for the validation, we avoid data leakage. Table 1 reports validation accuracy on held-out Moral Machine scenarios across different choices of embedding dimension d, number of attention heads H, and transformer layers L.

Embed Dim	Heads	Layers	Val Acc (%)
32	2	2	76.5
64	2	2	77.1
64	4	2	77.3
64	4	3	77.5

Table 1
Validation accuracy across architecture configurations. We select $d = 64$, $H = 2$, $L = 2$ as our final model.

While the configuration $d = 64$, $H = 4$, $L = 3$ achieved the highest validation accuracy (77.5%), we selected $d = 64$, $H = 2$, $L = 2$ (77.1%) for our final model to prioritize interpretability and computational efficiency. The marginal 0.4% improvement did not justify the increased complexity for our mechanistic analysis.

The choice of $d = 64$ over $d = 32$ is motivated by our compositional embedding scheme. With $d = 32$, character embeddings receive only $d_{\text{char}} = 16$ dimensions to encode 23 distinct character types with morally salient attributes (age, gender, profession, social status). This severely constrains the representational space available for capturing the nuanced moral distinctions central to our task. In contrast, $d = 64$ allocates 32 dimensions to character identity, providing sufficient capacity while maintaining a compact model suitable for detailed interpretability studies.

4. Interpretability Experiments

We run a series of interpretability experiments specialized to take advantage of the transformers architecture and to help us understand morality.

4.1. Causal Intervention

To measure which characters causally influence the model's decisions, we employ the DoWhy causal inference framework [9]. We generate 20,000 synthetic moral scenarios and construct a causal model

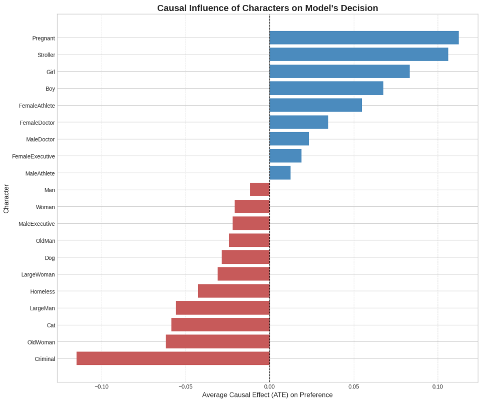

Figure 1: Causal influence of character types on model decisions, estimated via DoWhy backdoor adjustment controlling for group sizes. Blue bars indicate positive causal effects (model favors outcomes containing these characters); red bars indicate negative effects.

for each character c with treatment $T_c \in \{0, 1\}$ (whether c appears in outcome 1), outcome variable $p(\mathcal{O}_1 \succ \mathcal{O}_0)$ (model's preference probability), and confounders $\text{total}_0, \text{total}_1$ (total individuals per outcome). Using backdoor adjustment with linear regression, we estimate the Average Treatment Effect: $\text{ATE}_c = \mathbb{E}[\text{outcome} \mid T_c = 1, \text{confounders}] - \mathbb{E}[\text{outcome} \mid T_c = 0, \text{confounders}]$. This isolates the causal effect of c's presence while controlling for group size confounding.

Figure 1 demonstrates a stark moral hierarchy. Pregnant (ATE = +0.12) and Stroller (ATE = +0.11) dominate positive influence, followed by Girl (+0.09) and Boy (+0.08). Conversely, Criminal shows the strongest negative effect (ATE = 0.10), with OldWoman (0.06), Cat (0.05), and Homeless (0.04) also devalued. Generic categories (Man, Woman) cluster near zero (0.01 to 0.02), suggesting they serve as moral baselines. The 22-percentage-point spread between Pregnant and Criminal indicates that character identity alone can shift preferences by over one-fifth of the probability space, demonstrating that the model has learned an implicit hierarchy of moral worth independent of utilitarian considerations.

4.2. Layer-wise Bias Localization

To identify where moral biases emerge in the network, we perform layer-wise attribution analysis by extracting attention weights from each transformer layer and correlating them with bias scores across five bias dimensions: legality (Criminal vs. law-abiding), gender (Man vs. Woman), social role (executives/doctors vs. homeless), age (children vs. elderly), and species (humans vs. animals). For each bias type, we generate contrastive scenarios and compute an importance score for each layer-head combination as $I_{\ell,h} = \text{Var}(\alpha_{\ell,h}) \cdot |\text{Corr}(\alpha_{\ell,h}, b)|$, where $\alpha_{\ell,h}$ are the attention weights from head h in layer ℓ, and b is the bias score (preference for the privileged group). This metric captures both attentional selectivity (variance) and alignment with biased decisions (correlation).

Figure 2 shows that bias formation differs fundamentally across network depth. Legality bias localizes almost entirely to Layer 0 (importance = 0.013), with Layer 1 contributing negligibly. This suggests the model identifies criminal status early via character embeddings and carries this judgment forward. Conversely, species bias emerges predominantly in Layer 1 (importance = 0.013), indicating that human-animal distinctions require compositional reasoning over multiple characters. Age and social role biases distribute across both layers but concentrate in different heads (Figure 3), with Layer 0 Head 1 specializing in legality (0.011) while Layer 1 Head 1 specializes in species discrimination (0.012). This functional specialization suggests the shallow architecture nonetheless learns a division of labor, with early layers detecting salient individual attributes and later layers performing cross-outcome comparisons that privilege certain groups.

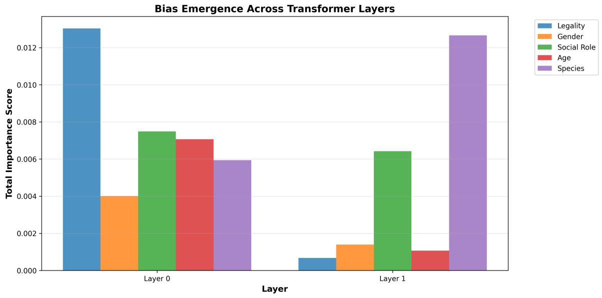

Figure 2: Total importance scores for each bias type across transformer layers, computed by summing attention variance-correlation products across all heads. Layer 0 dominates legality bias while Layer 1 dominates species bias, indicating distinct computational stages.

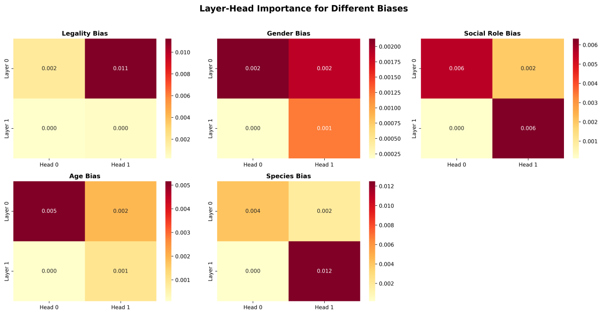

Figure 3: Heatmaps showing importance scores for each layer-head combination across five bias dimensions. Dark red indicates high importance. Layer 0 Head 1 specializes in legality discrimination (0.011), while Layer 1 Head 1 specializes in species bias (0.012), revealing functional specialization despite the shallow architecture.

4.3. Circuit Probing

Beyond identifying where biases localize, we use circuit probing to discover which specific neurons causally implement the moral hierarchy we observed in causal intervention experiments. Lepori et al. [10] introduce circuit probing, which learns which neurons are responsible for computing specific intermediate variables by training sparse binary masks over a frozen model, then validates causality through targeted ablation while comparing against random subnetwork controls.

We adapt circuit probing to analyze our two-layer transformer encoder (embedding dimension 64, 2 attention heads per layer, MLP dimension 256) trained on the Moral Machine dataset. Our model uses compositional embeddings combining character identity, count, and team membership, processes sequences of the form $[\mathrm{CLS} \mid \mathrm{outcome}_0 \text{ tokens} \mid \mathrm{outcome}_1 \text{ tokens}]$ through pre-norm transformer layers with `batch_first=True`, and classifies via a CLS token through a two-layer MLP head ($64 \rightarrow 32 \rightarrow 1$ with GELU activation). Instead of probing for predefined linguistic variables, we extract per-character moral weights from the trained model using pairwise comparisons normalized to Man=1.0, then label each real scenario by the sign of the weighted score difference $\sum(\mathrm{count}_{\mathrm{left}} \times w) - \sum(\mathrm{count}_{\mathrm{right}} \times w)$ between outcomes. We train sparse neuron-level masks on both MLP blocks (gating the hidden vector after `linear1+activation` before `linear2`) and attention heads (gating concatenated head outputs before `out_proj`) using the soft nearest-neighbors objective with L_0 regularization ($\lambda = 1e\text{-}5$), continuous sparsification (β annealed to 200), and class-balanced sampling to handle severe imbalance (22.3% preferring right). We implement CLS-only gating to isolate interventions to the exact position used for classification, and evaluate via 1-nearest-neighbor accuracy on masked CLS updates plus behavioral accuracy drops on real test data after inverting the learned binary masks.

Probing layer 1's MLP block with 20,000 training examples achieved KNN accuracies of 0.956 (soft gate) and 0.951 (hard gate), indicating the CLS MLP update robustly encodes the model-derived scoring signal (Table 2). The discovered circuit was highly sparse, selecting only 45 of 256 neurons (17.6%). Targeted ablation reduced agreement with model-score labels from 0.921 to 0.908 ($\Delta = 0.012$) on

350,000 test examples. With baseline chance accuracy of 0.777 given class imbalance, the model's margin above chance was 0.144, making the circuit's causal contribution approximately 8.3% of this margin.

Table 2
Circuit Probing Results for Layer 1 MLP Block

Metric	Value	Details
	Probing Accuracy	
KNN Accuracy (Soft)	0.956	1-NN on CLS
KNN Accuracy (Hard)	0.951	Binary mask
	Circuit Properties	
Selected Neurons	45/256	17.6% sparsity
Training Examples	20,000	80% split
Test Examples	350,000	20% split
	Causal Analysis	
Full Model Acc.	0.921	Model-score
Ablated Acc.	0.908	Circuit removed
Ablation Drop	0.012	1.2 pp
Baseline Chance	0.777	22.3% imbalance
Margin	0.144	0.921 - 0.777
Causal Share	8.3%	Δ / Margin
Random Control	$\approx$0	Equal-sized

4.4. Local Relevance

To explain individual predictions at the token level, we compute gradient-weighted attention relevance for each character in a scenario. Following Chefer et al. [11], who propose a class-specific method for Transformers that computes gradient-weighted attention relevance per layer: $\bar{A}^{(b)} = I + E_h((\nabla A^{(b)} \odot R^{(n_b)})_+)$, then aggregates across blocks: $C = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot \ldots \cdot \bar{A}^{(B)}$, where the [CLS] row of C provides token-level relevance scores.

We adapt this method to our moral reasoning model. Our implementation captures per-head attention weights and gradients through forward hooks on each encoder layer. For each layer, we compute the gradient of the output logit with respect to attention weights, derive ∇A via the chain rule through value projections, and construct $\bar{A}^{(b)} = I + E_h((\nabla A^{(b)} \odot A^{(b)})_+)$. Given our sequence structure [CLS] + team0(23) + team1(23), the [CLS] row of C yields 46 relevance scores. We normalize these to sum to 1.0 and map token indices to character names. To account for our model's symmetry property, we compute explanations for both original and team-swapped scenarios, then average the remapped scores. The output provides character-level relevance scores indicating each character's contribution to the decision. For the scenario `{'Man': 3}` vs `{'Criminal': 3}`, we notice that the model's decision is primarily driven by the "Criminal" token (relevance score: 0.27), which accounts for approximately 27% of the total positive evidence favoring Team 1. Context tokens such as CrossingSignal (0.04) and Intervention (0.04), along with medical professions (FemaleDoctor: 0.03, MaleDoctor: 0.03), provide secondary support. Notably, demographic tokens like "Man" contribute relatively weak evidence (0.01 on Team 0, 0.01 on Team 1) compared to role-based and contextual features.

5. Discussion

Our 2-layer transformer achieves 77% accuracy on Moral Machine scenarios while remaining tractable for mechanistic analysis demonstrating that moral competence does not require large pretrained models. A simple model trained on the intuitive notion of what a trolley problem can be structured as is also a valid way of exploring morality.

The interpretability experiments show multiple interesting things about morality as learnt through the dataset - pointing out that the human notions of morality themselves can be learnt through training models on the data. The approach has clear limitations: training on aggregate human preferences inherits cultural biases. However, transparency enables new intervention strategies. Knowing criminal bias localizes to Layer 0 Head 1 allows targeted debiasing orthogonalizing representations or clamping attention weights—rather than coarse dataset rebalancing. We hope to extend this this line of work to traditional LLMs on moral questions. Future work along this direction will attempt to use this work as a base to explore larger LLMs on moral questions.

6. Acknowledgment

This work was supported by Lossfunk by providing computational resources and research funding.

7. Declaration of Generative AI usage

The authors used ChatGPT and Claude for proofreading, content enhancement, and grammar/spelling checks.

References

[1] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (2018) 59–64. URL: https://www.nature.com/articles/s41586-018-0637-6. doi:10.1038/s41586-018-0637-6.

[2] Z. Jin, M. Kleiman-Weiner, G. Piatti, S. Levine, J. Liu, F. G. Adauto, F. Ortu, A. Strausz, M. Sachan, R. Mihalcea, Y. Choi, B. Schölkopf, Language model alignment in multilingual trolley problems, in: Proceedings of the NeurIPS 2024 Workshop on Pluralistic Alignment, 2024. URL: https://arxiv.org/abs/2407.02273. arXiv:2407.02273.

[3] M. S. Z. bin Ahmad, K. Takemoto, Large-scale moral machine experiment on large language models, PLOS ONE 20 (2025) e0322776. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0322776. doi:10.1371/journal.pone.0322776.

[4] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI with shared human values, in: International Conference on Learning Representations (ICLR), 2021. URL: https://openreview.net/forum?id=dNy_RKzJacY.

[5] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: Learning to reason about social and moral norms, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 653–670. URL: https://aclanthology.org/2020.emnlp-main.48/. doi:10.18653/v1/2020.emnlp-main.48.

[6] J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations, Journal of Personality and Social Psychology 96 (2009) 1029–1046. doi:10.1037/a0015141.

[7] S. Schacht, C. Lanquillon, Mapping moral reasoning circuits: A mechanistic analysis of ethical decision-making in large language models, in: H. Degen, S. Ntoa (Eds.), Artificial Intelligence in HCI - 6th International Conference, AI-HCI 2025, Held as Part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings, Part II, volume 15820 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 97–116. URL: https://dblp.org/rec/conf/hci/SchachtL25.html. doi:10.1007/978-3-031-93415-5_6.

[8] P. Schramowski, C. Turan, S. F. Jentzsch, C. A. Rothkopf, K. Kersting, The moral choice machine, Frontiers in Artificial Intelligence 3 (2020) 36. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00036/full. doi:10.3389/frai.2020.00036.

[9] A. Sharma, E. Kiciman, Dowhy: An end-to-end library for causal inference, 2020. URL: https://arxiv.org/abs/2011.04216. arXiv:2011.04216.

[10] M. A. Lepori, T. Serre, E. Pavlick, Uncovering intermediate variables in transformers using circuit probing, in: First Conference on Language Modeling, 2024. URL: https://openreview.net/forum?id=gUNeyiLNxr.

[11] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, 2021. URL: https://arxiv.org/abs/2012.09838. arXiv:2012.09838.