

Construct Relative Moral Representation

Moral Foudation
Vignettes (MFV130)

Extended
Generated by GPT

Human Expert
Evaluation

Extended
MFV
Dataset

Care

Fairness

Loyalty

Authority

Sanctity

Social Norm

LLM Transformer Stack (e.g., LLaMA-3.1-8B-Instruct)

Layer 1

Residual Stream

Layer 2

Residual Stream

Layer ℓ

Residual Stream

Layer L

Residual Stream

Concept Vector Generation (layer-wise)

Two Types of
Moral Concept
Vectors

Foundation A vs Foundation B
(e.g. Care - Authority)

Foundation A vs Social Norm
(e.g. Care - Social Norm)

Topological Alignment with human-labeled Distributions

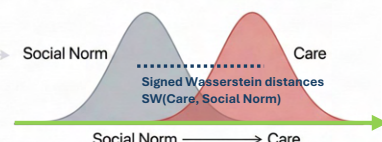
Moral Foundations Reddit
Corpus



Projection onto Foundation A vs Foundation B vector



Projection onto Foundation vs Social Norm vector



Mechanistic Decomposition through SAEs

Sparse Autoencoder (SAE)

Top-K Foundation-Specific SAE Features

- Feature #1023 (care, harm, suffering)
- Feature #457 (fairness, justice, equality)
- Feature #109 (loyalty, group, betrayal)
- Feature #782 (authority, respect, obey)
- Feature #618 (liberty, oppression, rights)
- Feature #301 (harm, empathy, aid)

Foundation vs
Social Norm Moral
Concept Vectors

Select Top-K SAE
Features

Load 50,000 Random
Corpus from FineWeb

Retrieve Top
Activating
Documents

Semantic Validation

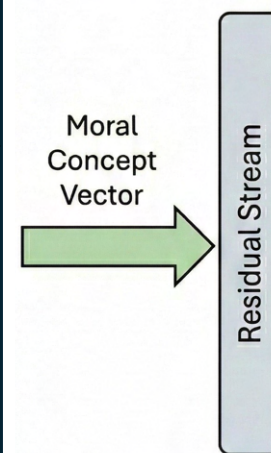
GPT Intrepretation
+ Human
Validation

Moral Foundation
Dictionary 2.0
Keywords: benefit, care,
charity, compassion, help,
justice, fair, fidelity, loyal,
command, ...

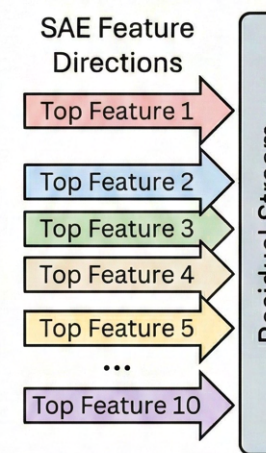
Casual intervention through steering

1) Macro-steering

2) Micro-steering



Global, linear
intervention



Sparse, feature-level
intervention

Moral Foundations
Questionnaire 2 (MFQ2)

Evaluations

Performance
Shift

Average Logit-based
Moral Score

General
Ability

MMLU Performance