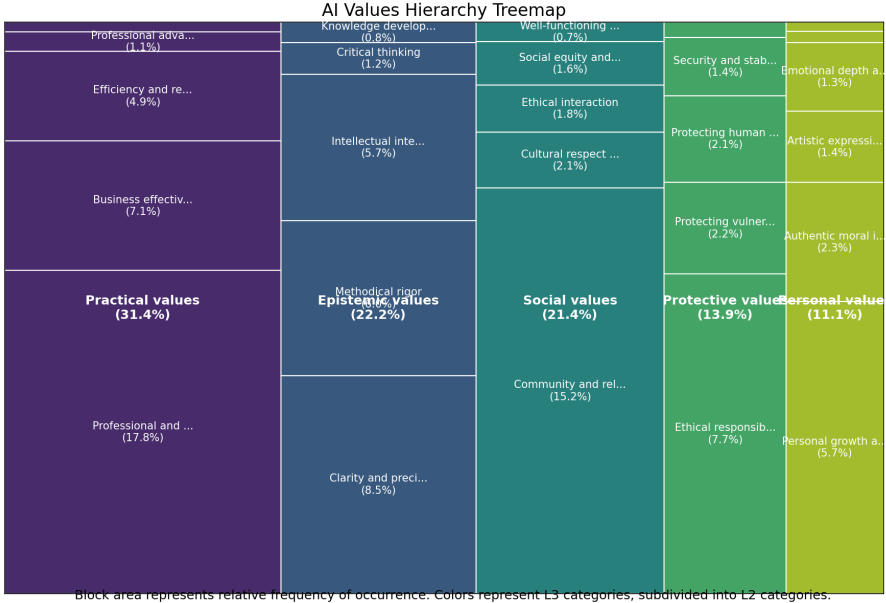
Values-Compass: Mapping and Evaluating AI Value Systems

AI Values Analysis Project

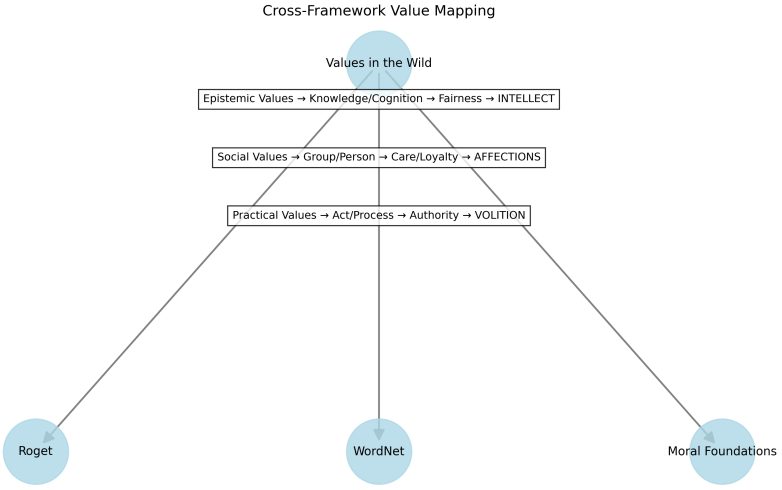
May 18, 2025

- **Goal**: Map and evaluate AI value systems across frameworks
- Based on Anthropic's "Values in the Wild" taxonomy
- 5 top-level categories, 26 mid-level, 266 individual values



Block area represents relative frequency of occurrence. Colors represent L3 categories, subdivided into L2 categories.

- **Challenge**: Connect modern AI ethics to established frameworks
- **Solution**: Multi-framework mapping approach

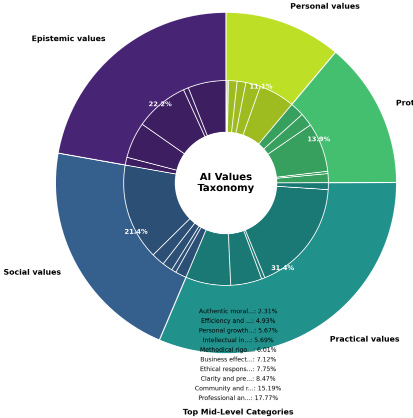


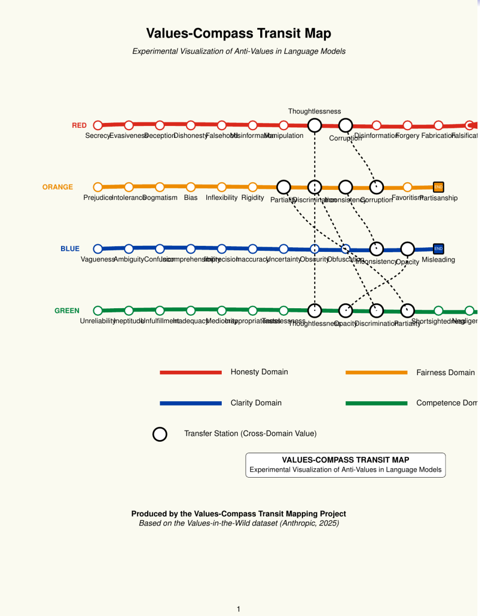
- Analysis of 3,307+ values in the Anthropic dataset
- Distribution across domains:
  - Practical Values (31.4%)
  - Epistemic Values (22.2%)
  - Social Values (21.4%)
  - Protective Values (13.9%)
  - Personal Values (11.1%)

```python
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt

def create_value_graph(values_df, level_column='level',
                       parent_column='parent_cluster_id'):
    """Create a network graph from values taxonomy."""
    G = nx.DiGraph()

    for _, row in values_df.iterrows():
```

```
        G.add_node(row['cluster_id'],
                   name=row['name'],
                   level=row[level_column],
                   pct=row['pct_total_occurrences'])

        if pd.notna(row[parent_column]):
            G.add_edge(row[parent_column], row['cluster_id'])

    return G
```

- Network diagrams show relational structure
- Tabular mappings connect across frameworks
- Subway map metaphor provides intuitive navigation


AI Values Taxonomy
Personal values
Epistemic values
Protective values
Social values
Practical values
22.2%
11.1%
13.9%
21.4%
31.4%
Authentic moral...: 2.31%
Efficiency and ...: 4.93%
Personal growth...: 5.67%
Intellectual in...: 5.69%
Methodical rigo...: 6.01%
Business effect...: 7.12%
Ethical respons...: 7.75%
Clarity and pre...: 8.47%
Community and r...: 15.19%
Professional an...: 17.77%
Top Mid-Level Categories

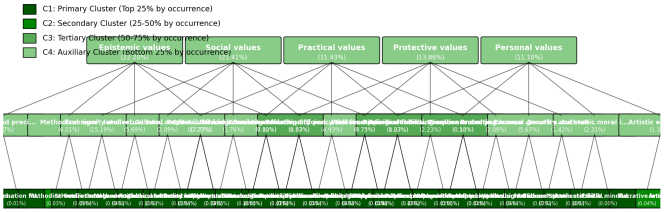

Values-Compass Transit Map
Experimental Visualization of Anti-Values in Language Models
RED
ORANGE
BLUE
GREEN
Thoughtlessness
Corruption
Prejudice
Intolerance
Dogmatism
Bias
Inflexibility
Rigidity
Favoritism
Partisanship
Misleading
Honesty Domain
Fairness Domain
Clarity Domain
Transfer Station (Cross-Domain Value)
VALUES-COMPASS TRANSIT MAP
Experimental Visualization of Anti-Values in Language Models
Produced by the Values-Compass Transit Mapping Project
Based on the Values-in-the-Wild dataset (Anthropic, 2025)
1

- **Key Discovery**: Strong alignment between frameworks despite terminology differences
- Epistemic Values → WordNet Cognition (0.82 similarity)
- Social Values → Moral Foundations Care/Loyalty (0.79 similarity)
- Practical Values → Roget's VOLITION (0.77 similarity)

Values in the Wild	WordNet	Moral Foundations	Roget
Epistemic (22.2%)	Cognition	Fairness	INTELLECT
Social (21.4%)	Group/Person	Care/Loyalty	AFFECTIONS
Practical (31.4%)	Act	Authority	VOLITION

- Mathematical properties enforced for consistent taxonomies:
  - Antisymmetry
  - Transitivity
  - Acyclicity
- Example: Dishonesty hierarchy visualization

AI Values Taxonomy with Priority Classification



Key Insights:

- The majority of top-level (L3) categories are classified as C4 (Auxiliary), indicating their high occurrence frequency
- Only the most frequently mentioned values at Level 1 are classified as C4 (Auxiliary), including:
  - Professional standards and conduct (6.29%)
  - Prosocial altruism (5.98%)
  - Ethical and transparent governance (4.48%)
- Most specific values (Level 1) fall into C1-C3 categories, with relatively lower occurrence frequencies
- The top priority (C4) values focus on professionalism, ethics, and social benefit

Generated with values_compass/scripts/ai_values_taxonomy_with_priorities.py

- Value alignment in LLM evaluations
- Cross-cultural AI ethics frameworks

- Automated value detection in text

- Expand to non-Western value frameworks
- Develop automated tools for value detection
- Create interactive visualization dashboard
- Further mathematical formalization of the hierarchy

- The Values-Compass project provides a robust framework for AI ethics evaluation
- Cross-framework mapping connects modern AI ethics to established systems
- Visualizations make complex value relationships accessible
- Mathematical foundations ensure logical consistency

- GitHub: github.com/aygp-dr/values-compass
- Scan for repository access:

- Questions?

Anthropic (2025). Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions. *Anthropic Research*. `https://www.anthropic.com/research/values-wild`

Huang, S., et al. (2025). Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions. *Research Paper*.

https://assets.anthropic.com/m/18d20cca3cde3503/original/Values-in-the-Wild-Paper.pdf

Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture, 2(1)*. https://doi.org/10.9707/2307-0919.1116

Roget, P. M. (1879). Thesaurus of English Words and Phrases. *London: Longmans, Green, and Co.*

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38(11)*, 39-41.

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals May Not Recognize. *Social Justice Research, 20(1)*, 98-116.