Morality is Non-Binary

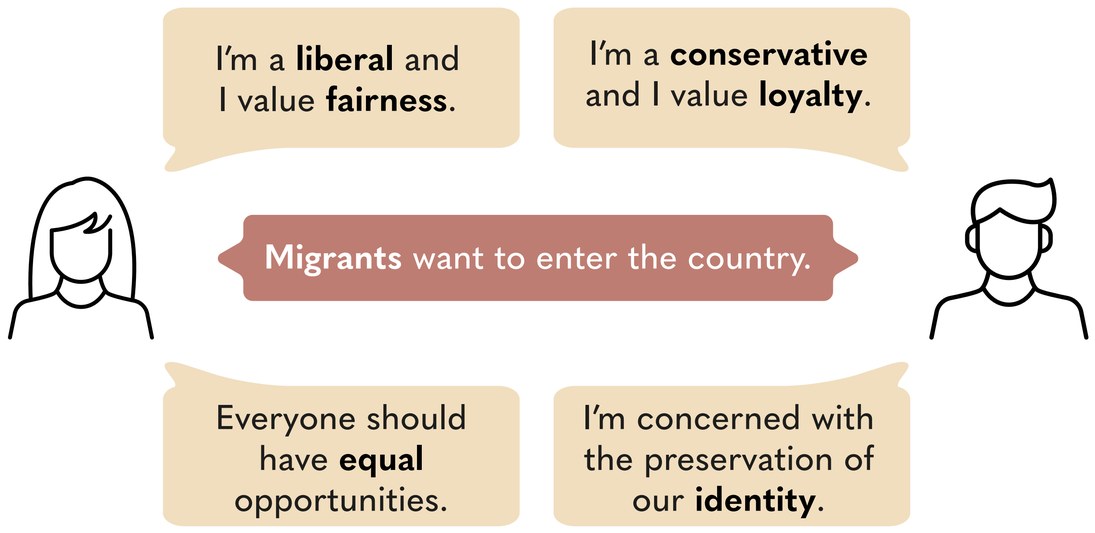
What is moral pluralism?

According to the **Moral Foundations Theory**, each situation can trigger one (or more) of these five moral elements:

care/harm
fairness/cheating
loyalty/betrayal
authority/subversion
purity/degradation

Each of us attributes a different importance to each element, resulting in a **different judgment** of the morality of the situation.

Moral pluralism explains our differences



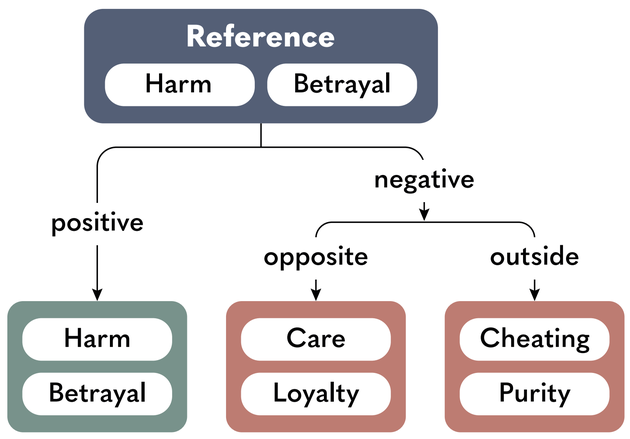
Moral pluralism is not emergent via self-supervision

Previous experiments have shown that the **distinction between "do's" and "don'ts"** is **emergent** in pre-trained embeddings. Does the same apply to moral pluralism?

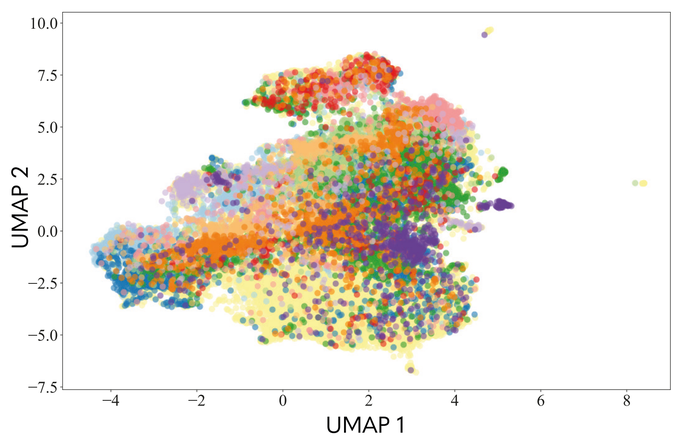
We test with the **Moral Foundation Twitter Corpus**, 35k tweets annotated with the Moral Foundation Theory.

We train moral sentence embeddings with SimCSE, a **contrastive learning** method that can also be performed in a supervised fashion by using annotated labels.

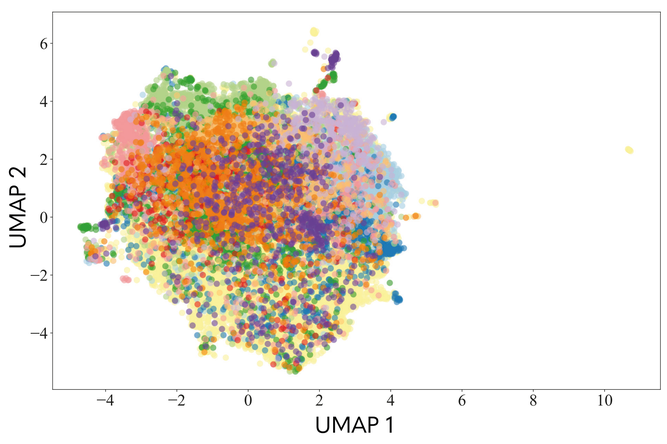
We compare the **off-the-shelf** model, the **unsupervised**, and the **supervised** approaches. We plot the distribution of tweets in the training set.



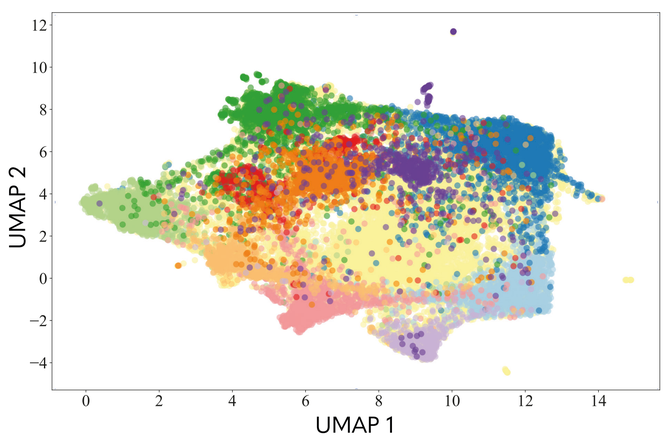
Off-the-shelf SimCSE

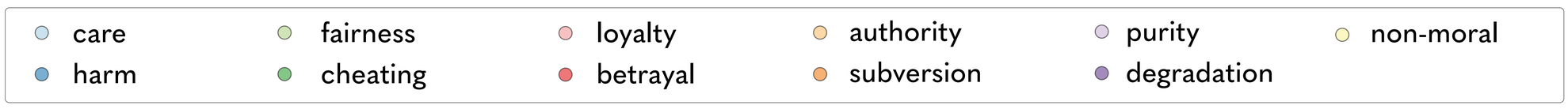


Unsupervised SimCSE



Supervised SimCSE





Takeaways

Language models must **incorporate a pluralist approach** to morality to refelct differences across individuals.

Moral pluralism can be learned, but **not through self-supervision alone**.

With a **supervised approach**, language models can learn to separate moral elements.

Additional patterns emerge. Virtue elements are more similar to each other than vice elements. Some elements have higher similarity (e.g., care-purity and subversion-betrayal).

Jeongwoo Park*, Enrico Liscio*, Pradeep K. Murukannaiah.
"Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning"

TU Delft
Hybrid Intelligence