Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Jeongwoo Park*, **Enrico Liscio***,
Pradeep K. Murukannaiah

TU Delft

Hybrid Intelligence

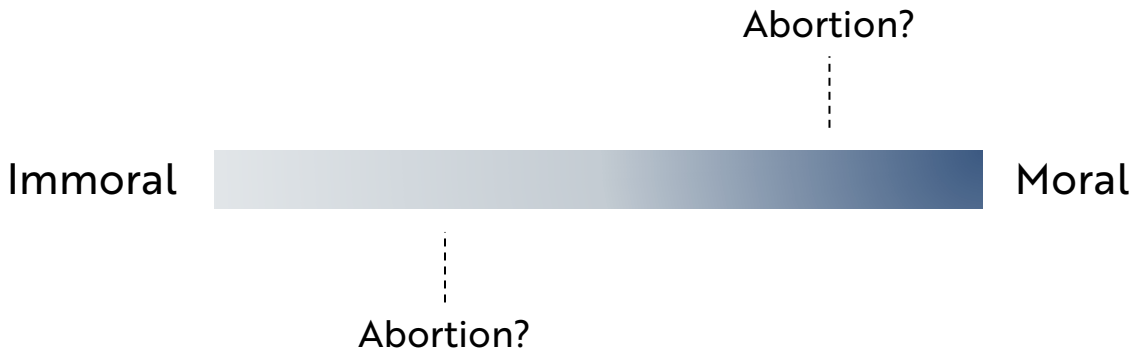
Human Morality

Morality helps us humans distinguish what is **right** from what is **wrong**.

In NLP, morality is often treated as a label on a **binary morality scale**. It has been shown that this binary approach to morality is emergent in language models.

Human Morality

Teaching language models an average perception of morality can lead to **dangerous biases**.



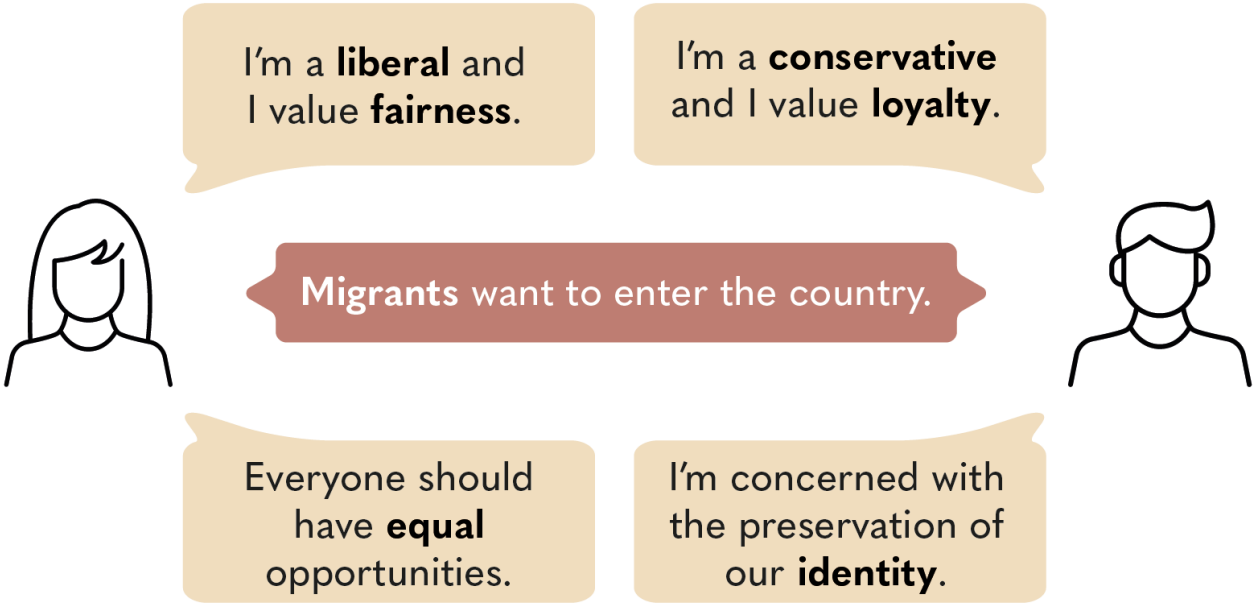
Moral Pluralism

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



Moral Pluralism

Is moral pluralism **emergent** in pre-trained unsupervised language models, or is a **supervised approach** necessary?

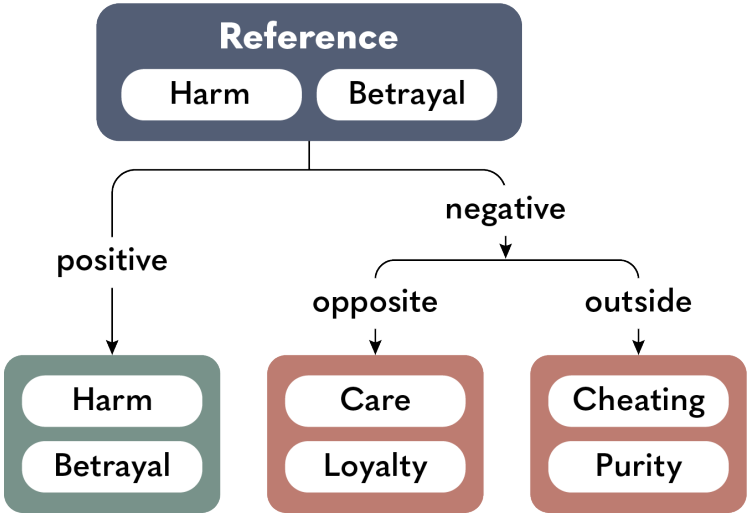
Experiments

We use **SimCSE**, a **contrastive learning** approach, to train BERT sentence embedding spaces with the **Moral Foundation Twitter Corpus** (35k tweets annotated with the Moral Foundation Theory).

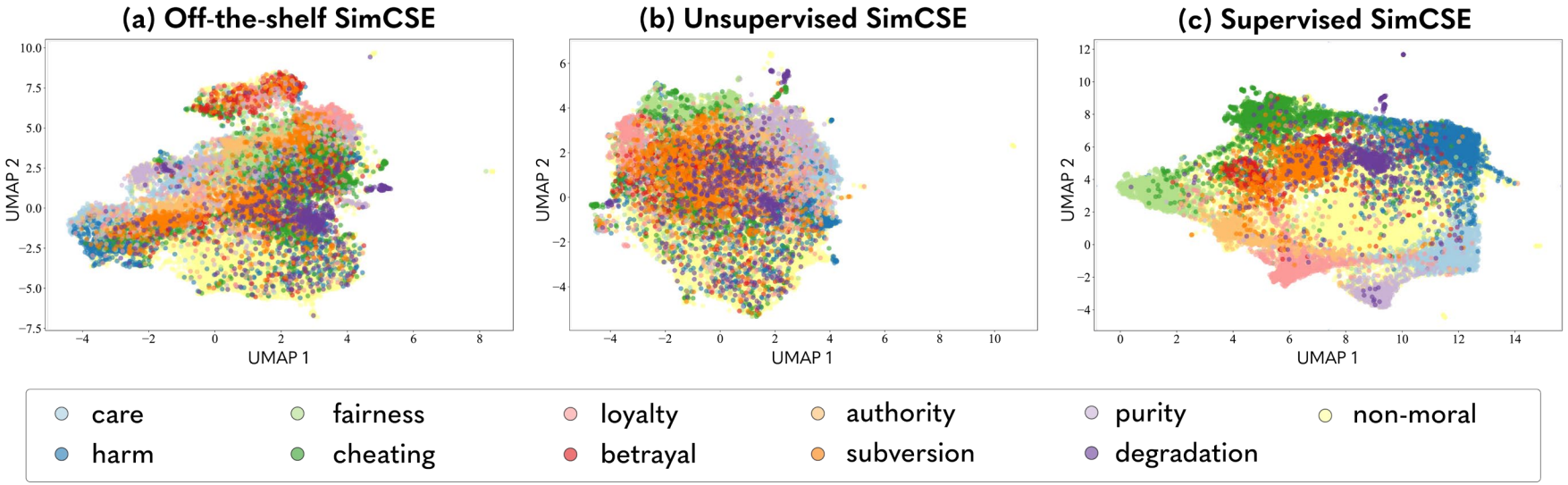
We compare the following approaches:

- **Off-the-shelf** model;
- **Unsupervised** training on the target dataset;
- **Supervised** training on the target dataset with the annotated labels.

Supervised Training



Results


(a) Off-the-shelf SimCSE
(b) Unsupervised SimCSE
(c) Supervised SimCSE
UMAP 1
UMAP 2
care
harm
fairness
cheating
loyalty
betrayal
authority
subversion
purity
degradation
non-moral

Takeaways

- Moral pluralism is **not emergent** via self-supervision alone but **can be learned** via a supervised approach.

- Interesting **patterns emerge**: virtue elements are more similar to each other than vice elements. Some elements have high similarity, e.g., care-purity and subversion-betrayal.

Thanks!

e.liscio@tudelft.nl
enricoliscio.github.io

Link to the paper!