What does a Text Classifier Learn about Morality?

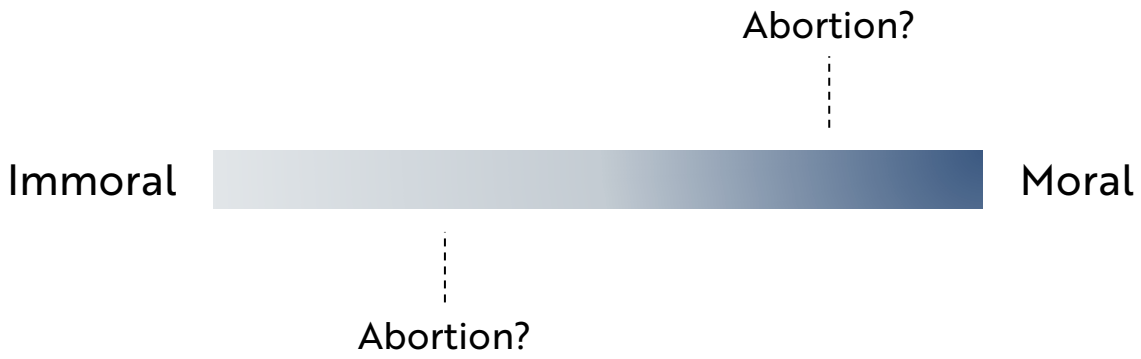
Enrico Liscio, Oscar Araque, Lorenzo Gatti,
Ionut Constantinescu, Catholijn M. Jonker,
Kyriaki Kalimeri, Pradeep K. Murukannaiah

Human Morality

Distinguishing what is **right** from what is **wrong**.

Human Morality in NLP

In NLP, morality is often treated as a label on a **morality scale**. However, teaching language models an average perception of morality leads to **dangerous biases**.



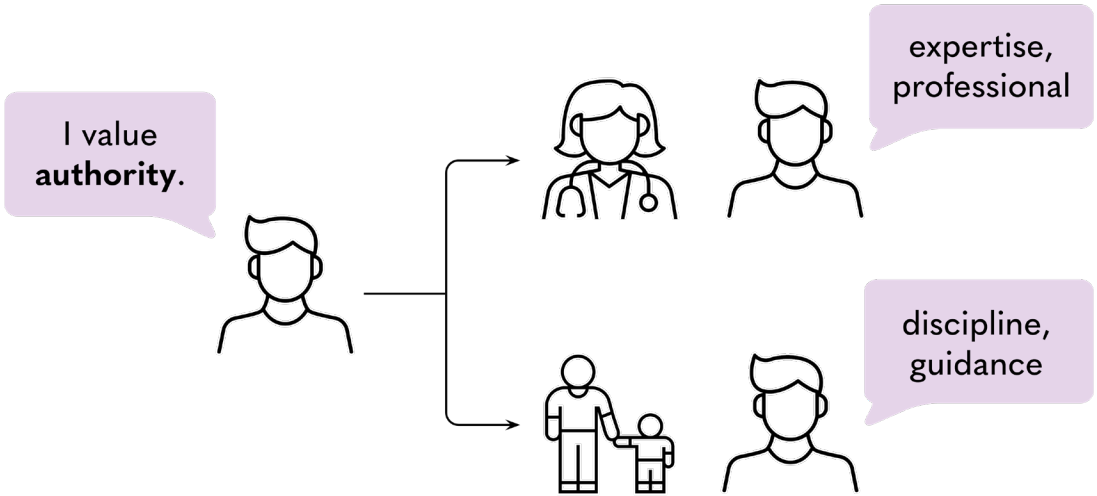
Moral Foundation Theory

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 Rhetoric is Domain-dependent



Do models detect domain-specific language?

Explaining Morality Classifiers

We propose **Tomea**, an XAI method for comparing morality classifiers across domains.

Tomea uses **SHAP** to generate lexicons of relevant words for each element in the Moral Foundation Theory. Then, it compares the generated lexicons **qualitatively** and **quantitatively** across domains, at moral element level and at domain level.

Experiments

Cross-domain comparison of BERT trained in the **seven** domains of the Moral Foundation Twitter Corpus (35k tweets):

#hatespeech
#Baltimoreprotests
#ALM
#BLM
#MeToo
#hurricaneSandy
#elections2016

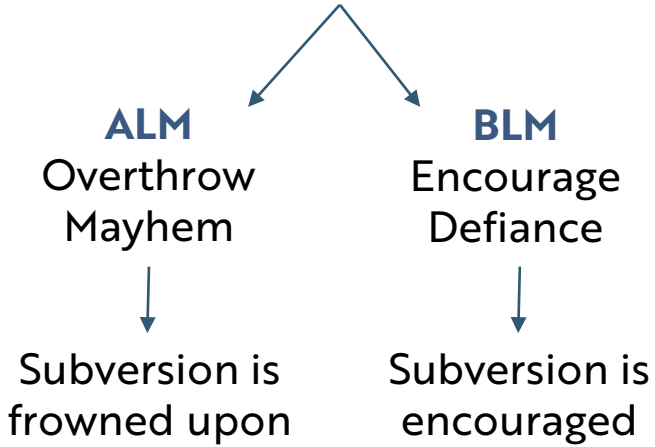
Quantitative Comparisons

Crowd workers moderately agree with the fine-grained moral lexicon similarities between domains (correlation of 0.4).

High Tomea similarity between domains entails better **out-of-domain performance** of the models (correlation of 0.79).

Qualitative Comparisons

ALM and **BLM** generally have a similar value rhetoric, but they differ for the element of *subversion*



Takeaways

- Language models recognize **small differences** in moral language in different domains.
- Practitioners must investigate the **qualitative similarity** between domains before using transfer learning.
- Small but **critical differences** between domains may not affect quantitative results but may **hinder usage** in a novel domain.