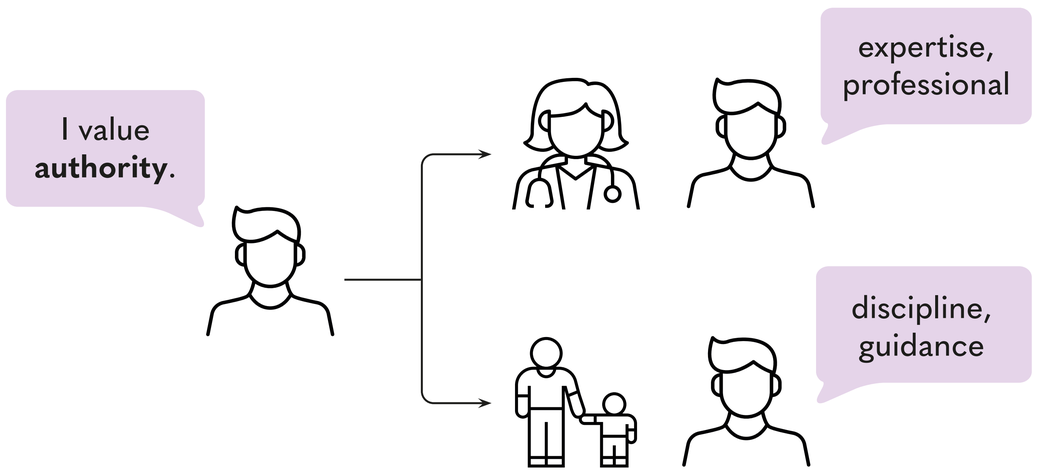
What does a Classifier Learn about Morality?

Moral values explain our differences

Moral rhetoric is domain dependent



Do models detect domain-specific language?

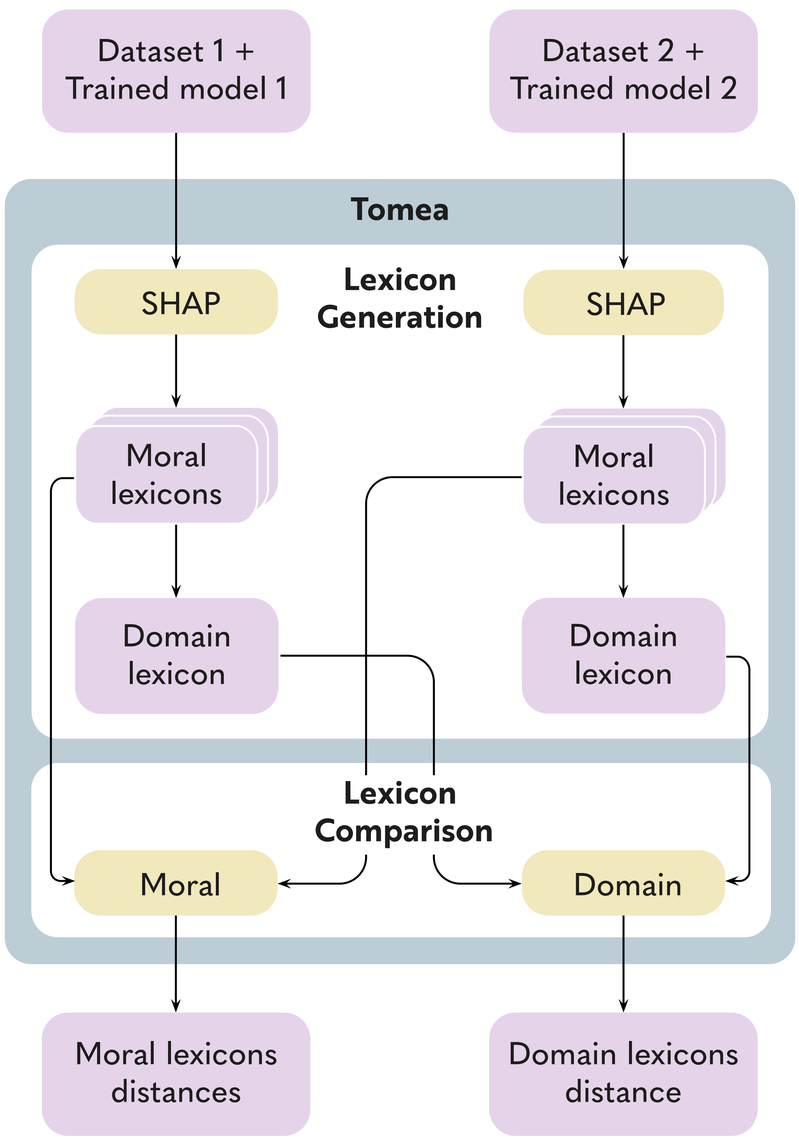
What is morality?

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.

Tomea: XAI method for comparing morality classifiers across domains



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

We perform **quantitative** and **qualitative** comparisons across domains.

Tomea's quantitative comparisons are reliable

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).

Tomea offers a qualitative moral rhetoric comparison

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

ALM
overthrow
mayhem
↓
Subversion is frowned upon

BLM
encourage
defiance
↓
Subversion is encouraged

Takeaways

Our experiments with Tomea show that language models recognize **small differences** in moral language in different domains.

Practicioners 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.

Tomea can be improved with rule-based and counterfactual **explanations**, and the qualitative analysis could be **systematized**.

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

TU Delft Hybrid Intelligence
Universidad Politécnica de Madrid
ETH zürich
University of Twente. ISI Foundation