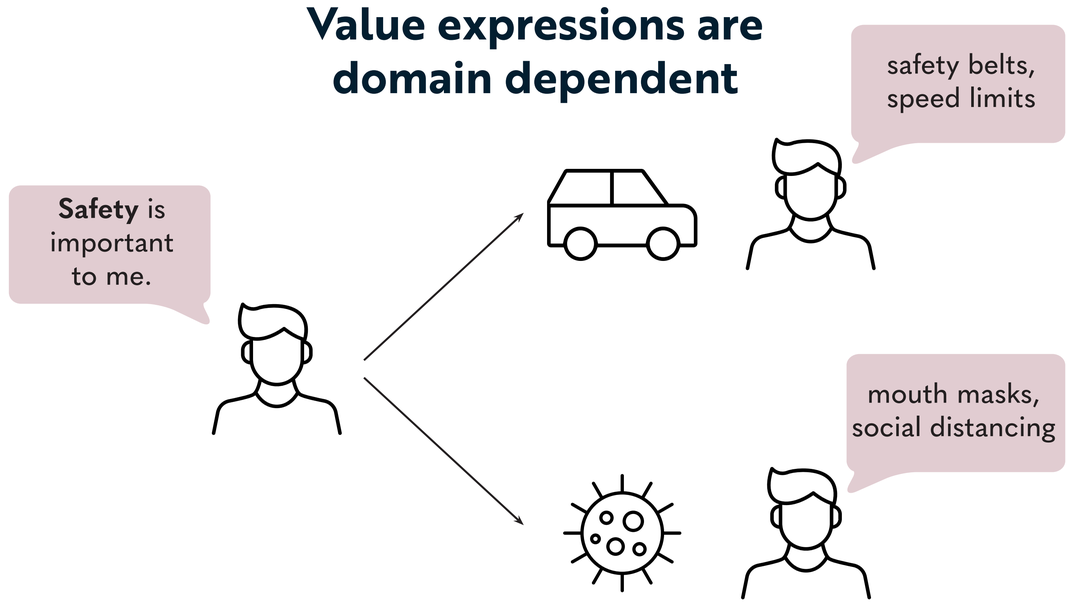
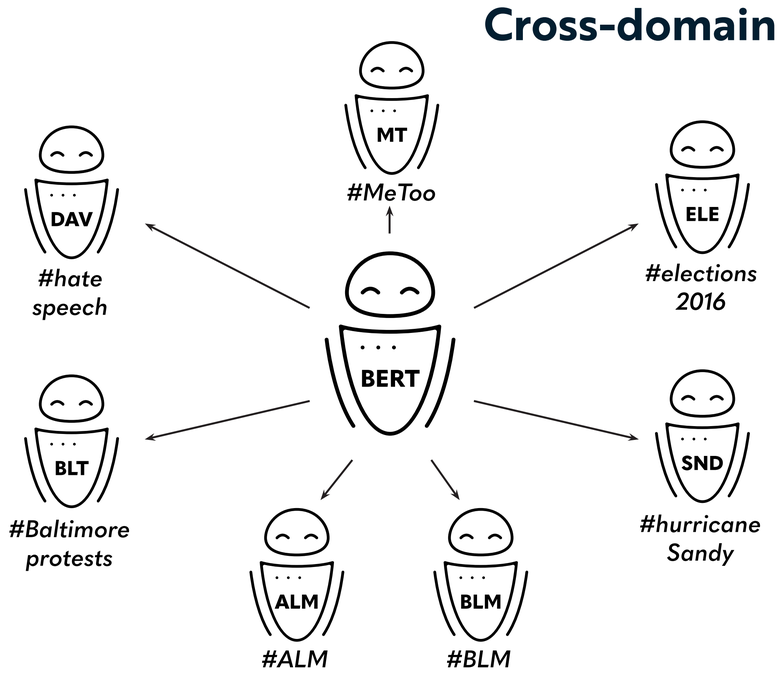
Cross-Domain Classification of Moral Values

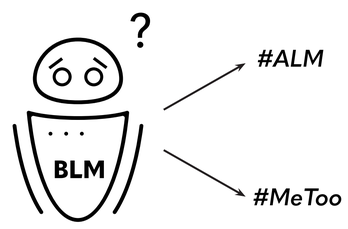
Values explain our differences

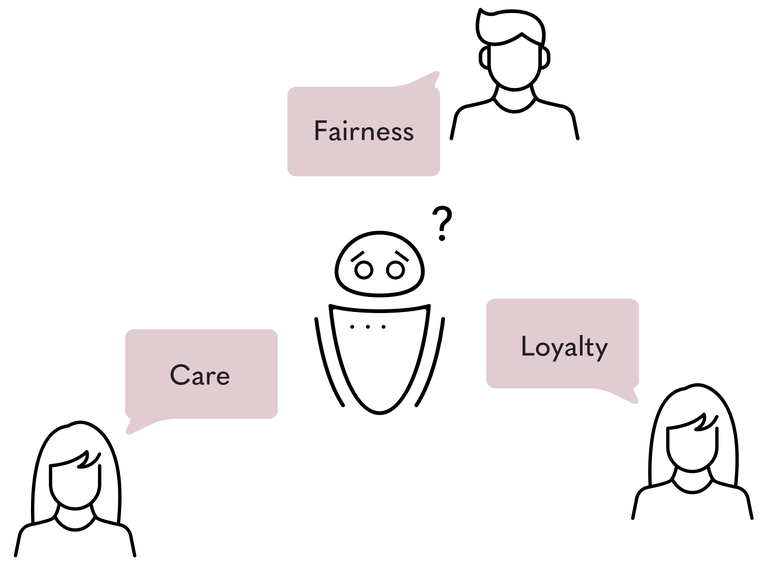
Value expressions are domain dependent



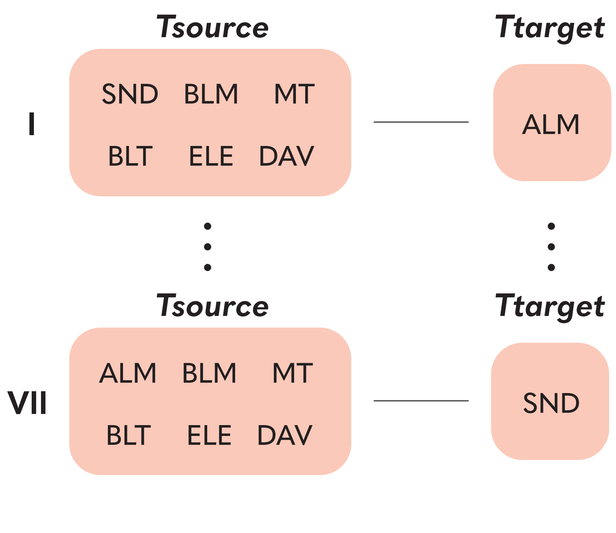
Cross-domain value classification with multiple annotators

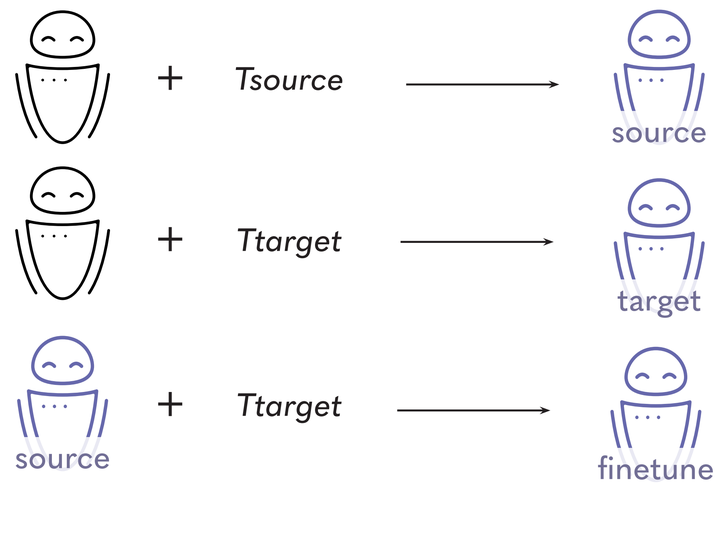






Experiments with datasets and training modalities combinations





Takeaways

A value classifier can **generalize** to novel domains, but its performance improves even when finetuned with a small portion of data.

Pretraining a value classifier yields **good performance** even when little training data is available.

Pretraining a value classifier yields **smaller confusion** among the moral values less frequent in the novel domain.

Catastrophic forgetting occurs even when finetuning on a small portion of data from the novel domain.

In the majority of classification errors, at least **one annotator agrees** with the model prediction.

We need to investigate methods for **incorporating annotators (dis-)agreement** in the model training.

Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, Pradeep K. Murukannaiah

TU Delft

Hybrid Intelligence