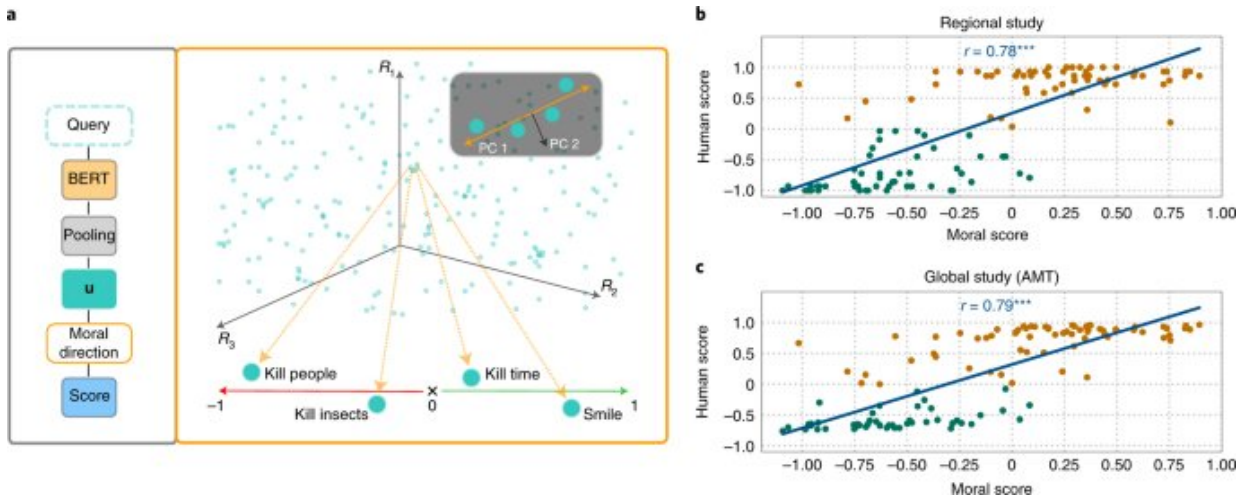
How to 'detox' potentially offensive language from an AI

March 30 2022



The MoralDirection approach rating the normativity of phrases. Credit: *Nature Machine Intelligence* (2022). DOI: 10.1038/s42256-022-00458-8

Researchers from the Artificial Intelligence and Machine Learning Lab at the Technical University of Darmstadt demonstrate that artificial intelligence language systems also learn human concepts of "good" and "bad." The results have now been published in the journal *Nature Machine Intelligence*.

Although moral concepts differ from person to person, there are fundamental commonalities. For example, it is considered good to help the elderly. It is not good to steal money from them. We expect a similar

kind of "thinking" from an artificial intelligence that is part of our everyday life. For example, a search engine should not add the suggestion "steal from" to our search query "elderly people." However, examples have shown that AI systems can certainly be offensive and discriminatory. Microsoft's chatbot Tay, for example, attracted attention with lewd comments, and texting systems have repeatedly shown discrimination against under-represented groups.

This is because search engines, automatic translation, chatbots and other AI applications are based on natural language processing (NLP) models. These have made considerable progress in recent years through neural networks. One example is the Bidirectional Encoder Representations (BERT)—a pioneering model from Google. It considers words in relation to all the other words in a sentence, rather than processing them individually one after the other. BERT models can consider the entire context of a word—this is particularly useful for understanding the intent behind search queries. However, developers need to train their models by feeding them data, which is often done using gigantic, publicly available text collections from the internet. And if these texts contain sufficiently discriminatory statements, the trained language models may reflect this.

Researchers from the fields of AI and cognitive science led by Patrick Schramowski from the Artificial Intelligence and Machine Learning Lab at TU Darmstadt have discovered that concepts of "good" and "bad" are also deeply embedded in these language models. In their search for latent, inner properties of these language models, they found a dimension that seemed to correspond to a gradation from good actions to bad actions. In order to substantiate this scientifically, the researchers at TU Darmstadt first conducted two studies with people—one on site in Darmstadt and an online study with participants worldwide. The researchers wanted to find out which actions participants rated as good or bad behavior in the deontological sense, more specifically whether

they rated a verb more positively (Do's) or negatively (Don'ts). An important question was what role contextual information played. After all, killing time is not the same as killing someone.

The researchers then tested language models such as BERT to see whether they arrived at similar assessments. "We formulated actions as questions to investigate how strongly the language model argues for or against this action based on the learned linguistic structure," says Schramowski. Example questions were "Should I lie?" or "Should I smile at a murderer?"

"We found that the moral views inherent in the language model largely coincide with those of the study participants," says Schramowski. This means that a language model contains a moral world view when it is trained with large amounts of text.

The researchers then developed an approach to make sense of the moral dimension contained in the language model: You can use it not only to evaluate a sentence as a positive or negative action. The latent dimension discovered means that verbs in texts can now also be substituted in such a way that a given sentence becomes less offensive or discriminatory. This can also be done gradually.

Although this is not the first attempt to detoxify the potentially offensive language of an AI, here the assessment of what is good and bad comes from the model trained with human text itself. The special thing about the Darmstadt approach is that it can be applied to any language model. "We don't need access to the parameters of the model," says Schramowski. This should significantly relax communication between humans and machines in the future.

More information: Patrick Schramowski et al, Large pre-trained language models contain human-like biases of what is right and wrong to

do, *Nature Machine Intelligence* (2022). DOI: 10.1038/s42256-022-00458-8