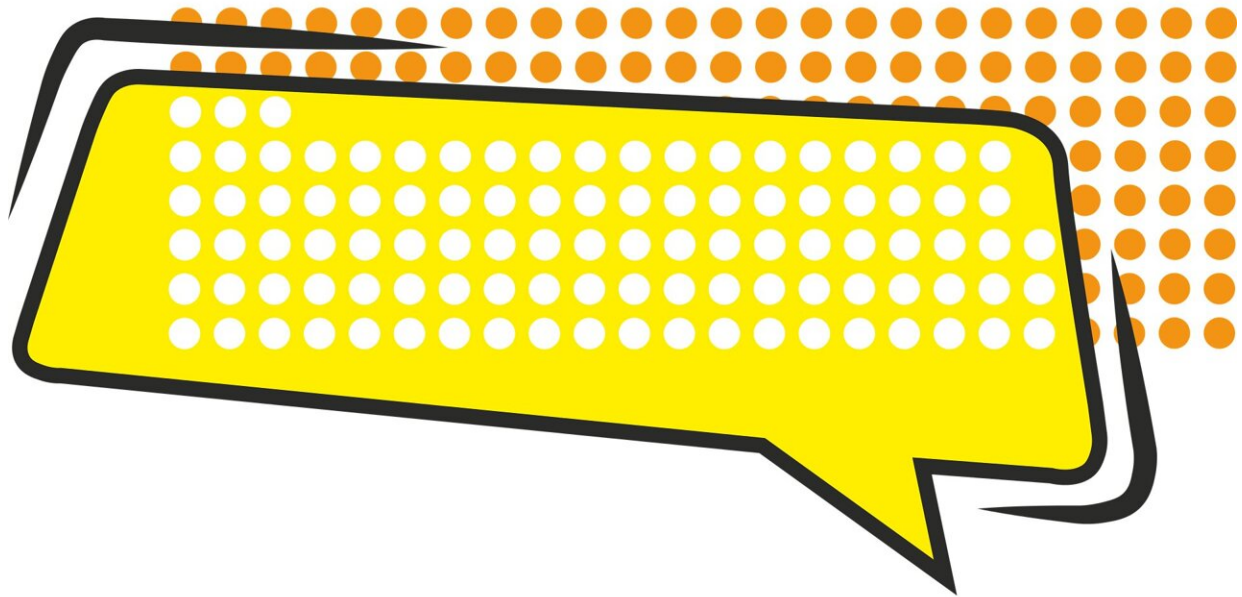# Exploring the moral foundations of hate speech

July 13 2023



Credit: Pixabay/CC0 Public Domain

Moral values such as purity and loyalty are often linked with hateful language, according to a study published in *PNAS Nexus*. Scholars in the field of natural language processing (NLP) have, in recent years, focused on improving the automated detection of hate speech so such language can be removed from online spaces. But some scholars have argued that mere detection is not a real solution—that NLP can and should be used to investigate the roots of hateful language.

Researcher Morteza Dehghani and colleagues have conducted three studies of hateful language, looking at speeches and texts written by leaders of the Nazi party between 1933 and 1945, hateful slurs in large text corpora across a multitude of languages, and [hate speech](#) from 2018 on the far-right social-media platform Gab. In each analysis, the study authors used NLP to situate hate [speech](#) in the context of moral motivations, worldviews, and rationales.

The authors used Moral Foundations Theory as their overarching framework. Moral Foundations Theory posits that moral sentiments can be classified into a small number of core values, including care, fairness, loyalty, authority, and purity. The authors found that Nazi propaganda appealed to the ideal of purity, in the sense that members of out-groups were depicted as impure and polluting.

In the large multi-lingual corpora, hateful [language](#) was highly associated with words related to the ideal of loyalty. On Gab, purity was again the most invoked value in hateful posts. Morality and hate may be deeply linked in the human psyche, according to the authors. Alternatively, or in addition, moral arguments may be used to legitimize hateful opinions.

provided for information purposes only.