The Meanings of Morality: Investigating the psychometric properties of distributed representations of latent moral concepts

Joe Hoover
University of Southern California, Los Angeles, California, United States

Katie Horton
University of Southern California

Morteza Dehghani
University of Southern California

Abstract: People's beliefs about what is morally right and wrong vary widely between individuals, contexts, and cultures; however it is thought that they are governed by core latent constructs. While there is evidence that these constructs are reflected in natural language, this requires further testing. We demonstrate that the structure of moral values in natural discourse can be modeled by applying factor analyses to distributed representations of morally relevant terms learned by a neural network. We first demonstrate that robust latent constructs can be estimated from the covariance of distributed representations of construct exemplars. We then test whether the factor structure proposed by Moral Foundations Theory (MFT) is reflected in natural language. Finally, we conduct a bottom-up investigation of the structure of moral values in natural language using free-responses reported by participants. Ultimately, we find evidence that the representation of moral values in natural language partially corresponds to MFT.