research highlights

COGNITIVE SCIENCE

Biased machines

Science **356**, 183–186 (2017)

GARY WATERS/IKON IMAGES/GETTY

Are implicit human biases reflected in the statistical properties of language? Research that combines machine-learning methods with insights from social psychology now suggests that this is indeed the case for a range of documented biases.

Aylin Caliskan from Princeton University and colleagues use the Implicit Association Test (IAT), which measures the strength of association between concepts (for example, black, female) and evaluations or stereotypes, as a source of human biases. They then develop a statistical analogue of the IAT and apply it to a corpus of text from the internet, where words are represented as vectors in multi-dimensional space based on the context in which each word occurs. Strength of association in the corpus is determined by measuring the distance between vectors. Using this method, the authors replicate in their 840-billion-word corpus a range of well-established IAT findings, including that musical instruments are much more pleasant than weapons; European-American names, as compared with African-American names, are significantly more associated with pleasant than unpleasant terms; and female names are more associated with family than career, when compared with male names.

The tool could be used to identify unknown implicit biases in texts and could provide invaluable insight into the minds and attitudes of historic populations that are unavailable for testing. It also shows, however, that without explicit coding of what is and is not societally acceptable, intelligent machines that learn from human text will also replicate any stereotypes embedded in that text. □

Stavroula Kousta