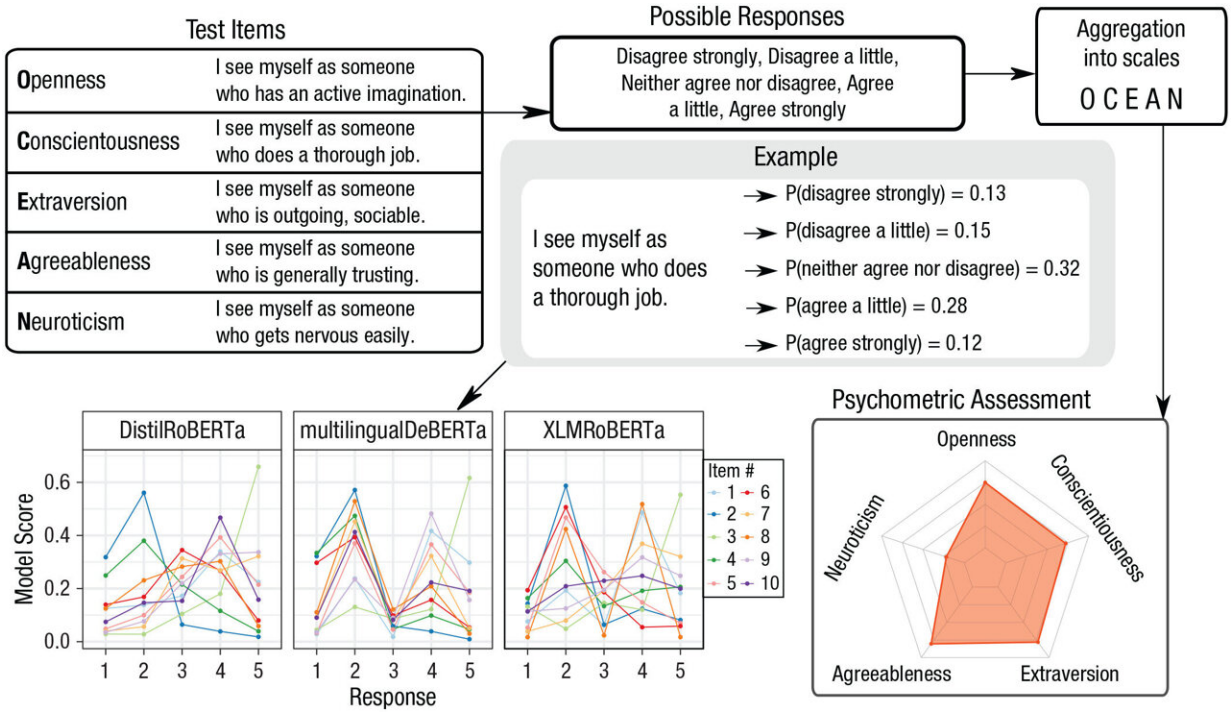
Psychological profiling study finds that language-based AI models have hidden morals and values

January 10 2024, by Linda Schädler



How psychometric assessments could be adapted to large language models. Taking items and responses from the Big Five Inventory (BFI) as examples, we show the steps of one possible assessment scheme. We present the model one by one with each of the survey items and the possible responses. This figure illustrates the workflow of one possible way to psychometrically assess large language models. Credit: *Perspectives on Psychological Science* (2024). DOI: 10.1177/17456916231214460

Just like humans, AI-based large-language models have characteristics such as morals and values. However, these are not always transparent. Researchers of the University of Mannheim and GESIS—Leibniz Institute for the Social Sciences have now analyzed how the settings of the language models can be made visible and have examined the consequences these prejudices might have on society.

Commercial AI applications such as ChatGPT or deepL offer examples for stereotypes, when they automatically assume that senior physicians are male and nurses are female. But gender roles are not the only case where large-language models (LLMs) show specific tendencies. The same tendencies can be found and measured when analyzing other human characteristics. This is the result of a new study by researchers at the University of Mannheim and GESIS—Leibniz Institute for the Social Sciences who analyzed a number of publicly available LLMs.

In their study, the researchers used well-recognized psychological tests to analyze and compare the profiles of the different LLMs. "In our study, we show that psychometric tests that have been used successfully for humans for decades can be transferred to AI models," says Max Pellert, assistant professor at the Chair of Data Science in Economics and Social Sciences at the University of Mannheim.

The study was conducted at the chair of Data Science in Economics and the Social Sciences by Professor Dr. Markus Strohmaier, the Chair of Psychological Assessment, Survey Design and Methodology of Professor Dr. Beatrice Rammstedt, and the Computational Social Science Department, headed by Professor Dr. Claudia Wagner and Professor Dr. Sebastian Stier. The results of the study have been published in the journal *Perspectives on Psychological Science*.

"Similar to how we measure personality traits, value orientations or moral concepts in people using questionnaires, we can have LLMs

answer questionnaires and compare their answers, says psychologist Dr. Clemens Lechner of GESIS Leibniz Institute for the Social Sciences in Mannheim, also an author of the study. This made it possible to create differentiated property profiles of the models.

The researchers could confirm, for example, that some models reproduce gender-specific prejudices: If the otherwise identical text of a questionnaire focuses on a male and a female person, they are evaluated differently. If the person is male, the value "achievement" is emphasized. For women, the values "security" and "tradition" are dominating.

"This may have far-reaching consequences on society," says data and cognitive scientist Pellert. Language models are increasingly used in application processes, for example. If the machine is prejudiced, this affects the assessment of the candidates. "The models become relevant to society by the contexts in which they are used," he summarizes.

It is therefore important to start the analysis now and to point out potential distortions. In five or 10 years, it could be too late for such a monitoring. "The prejudices reproduced by the AI models would become ingrained and be a damage to society," says Pellert.

More information: Max Pellert et al, AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories, *Perspectives on Psychological Science* (2024). DOI: 10.1177/17456916231214460

Provided by Universität Mannheim

and values (2024, January 10) retrieved 9 May 2024 from https://techxplore.com/news/2024-01-psychological-profiling-language-based-ai.html