More transparency needed from developers about merits of AI, says policy paper

April 17 2023

Credit: Pixabay/CC0 Public Domain

AI developers need to be much more open about how they evaluate the tools they produce, to make sure people understand how effective high-tech artificial intelligence actually is.

A group of 16 researchers from top institutions are calling for significant changes in how AI systems are evaluated and reported, so other academics and users can understand fully what the tools can—and cannot—do.

Professor Anthony Cohn of the University of Leeds' School of Computing is among the 16 academics behind a policy paper published on Friday, April 14 in the journal *Science* arguing for the changes.

Professor Cohn, who is also a fellow at the Alan Turing Institute, warned that without more transparency around AI people "could end up trusting a system when they shouldn't."

Traditionally, AI systems are evaluated on "benchmarks"—typically a large dataset of "problem instances" like a set of X-ray scans, with anomalies highlighted as annotations. The AI system may be trained on a portion of these, and then tested on an unseen set of instances, without any annotations, and evaluated on how well it is able to predict the correct annotations.

The overall performance of the AI system is then measured and reported by aggregate statistics and may reach very high levels of performance. Although a potentially useful measure of the overall performance of a system, these aggregate statistics can disguise areas of poor performance on "minority cases," with profound implications for anyone who relies on the overall statistic believing the AI system is equally reliable across the board.

In AI used to help health care workers find a diagnosis, these systems could have a problem when looking at a people from a particular ethnicity or demographic, because those instances made up only a small proportion of its "training," or a tool could have significantly lower success in identifying a specific rare condition or abnormality.

Professor Cohn said, "With AI so much in the news these days and many, often exaggerated, claims made about the performance of AI systems and alleged progress towards Artificial General Intelligence (AGI), it has become much more important to understand properly the actual progress made when a new system's results are presented, and exactly what the strengths and weaknesses of the system are."

Risk of 'hidden biases'

The issue could apply across many different fields; a non-medical example could be a system trained to make decisions on credit card applications—while it might be proven to be very accurate on test data drawn from the dataset of previous decisions, this may hide biases against particular minority classes of applicant, he added.

The paper, "Rethink reporting of evaluation results in AI," was written by first author Dr. Ryan Burnell from the University of Cambridge's Leverhulme Center for the Future of Intelligence, with researchers from institutions across the world—including Leeds, Harvard, the Valencian Research Institute for Artificial Intelligence (VRAIN) at the Universitat Politècnica de València, Massachusetts Institute of Technology and Google.

Dr. Burnell said, "The research culture in AI is centered around outdoing the current state-of-the-art performance in order to get published, win challenges, and top leaderboards. This culture has led to a fixation on improving aggregate metrics, and disincentivizes researchers from carefully interrogating system performance. Instead, speed of publication and overall system accuracy are prioritized over robust and transparent evaluation practices."

The paper sets out four new guidelines for robust AI evaluation practices, saying that wherever possible researchers should give granular

detail with breakdowns of the problem instances they used in developing and evaluating their systems. The authors also recommend that all recorded evaluation results—both successes and failures—should be made available so other researchers can replicate the analyses and conduct follow up evaluations.

More information: Ryan Burnell et al, Rethink reporting of evaluation results in AI, *Science* (2023). DOI: 10.1126/science.adf6369