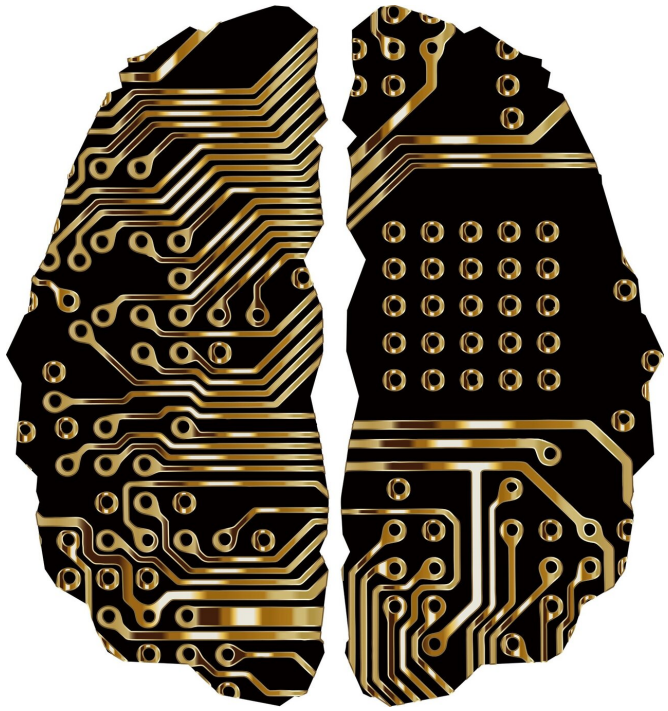
AI may mistake chess discussions as racist talk

18 February 2021



Credit: Pixabay/CC0 Public Domain

"The Queen's Gambit," the recent TV mini-series about a chess master, may have stirred increased interest in chess, but a word to the wise: social media talk about game-piece colors could lead to misunderstandings, at least for hate-speech detection software.

That's what a pair of Carnegie Mellon University researchers suspect happened to Antonio Radic, or "agadmator," a Croatian chess player who hosts a popular YouTube channel. Last June, his account was blocked for "harmful and dangerous" content.

YouTube never provided an explanation and reinstated the channel within 24 hours, said Ashiqur R. KhudaBukhsh a project scientist in CMU's Language Technologies Institute (LTI). It's nevertheless possible that "black vs. white" talk during Radi?'s interview with Grandmaster Hikaru Nakamura triggered software that automatically detects racist language, he suggested.

"We don't know what tools YouTube uses, but if they rely on artificial intelligence to detect racist language, this kind of accident can happen," KhudaBukhsh said. And if it happened publicly to someone as high-profile as Radic, it may well be happening quietly to lots of other people who are not so well known.

To see if this was feasible, KhudaBukhsh and Rupak Sarkar, an LTI course research engineer, tested two state-of-the-art speech classifiers—a type of AI software that can be trained to detect indications of hate speech. They used the classifiers to screen more than 680,000 comments gathered from five popular chess-focused YouTube channels.

They then randomly sampled 1,000 comments that at least one of the classifiers had flagged as hate speech. When they manually reviewed those comments, they found that the vast majority—82%—did not include hate speech. Words such as black, white, attack and threat seemed to be triggers, they said.

As with other AI programs that depend on machine learning, these classifiers are trained with large numbers of examples and their accuracy can vary depending on the set of examples used.

For instance, KhudaBukhsh recalled an exercise he encountered as a student, in which the goal was to identify "lazy dogs" and "active dogs" in a set of photos. Many of the training photos of active dogs showed broad expanses of grass because running dogs often were in the distance. As a result, the program sometimes identified photos containing large amounts of grass as examples of active dogs,

even if the photos didn't include any dogs.

In the case of chess, many of the training data sets likely include few examples of chess talk, leading to misclassification, he noted.

The research paper by KhudaBukhsh and Sarkar, a recent graduate of Kalyani Government Engineering College in India, won the Best Student Abstract Three-Minute Presentation this month at the Association for the Advancement of AI annual conference.

Provided by Carnegie Mellon University