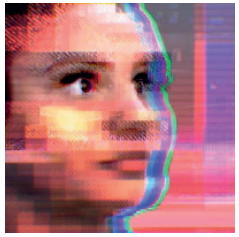
Bad bot! Bad!

Sometimes machines learn the wrong things



Tay

On March 23, 2016, Microsoft unveiled an artificially intelligent Twitter chatbot called Tay. Based on an existing chatbot that Microsoft had already been using in China, called Xiaoice, Tay was supposed to have the personality of a 19-year-old girl and interact with other Twitter users.

Unfortunately, some users began to realise that they could "train" Tay to say certain things, and the shenanigans began. Within 24 hours, Tay was firing off racist, xenophobic tweets. In one tweet, she screamed for a "RACE WAR". In another, she said that the Holocaust was "made up".

Microsoft shut her down after a day and apologised: "AI systems feed off both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical. We will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes."

Tay had a brief revival about a week after her short-lived debut, but was quickly shut down again after she began to spam meaningless messages and blamed her actions on alcohol.

Compas

Most US states rely on algorithmic risk-assessment systems that try to predict future behaviour of individual defendants and detainees. Those assessments are mostly used to guide bail conditions and sentencing terms.

One such system, called COMPAS, came under scrutiny in 2016 when ProPublica analysed its results in Broward County in Florida. What the news organisation found was that COMPAS was particularly likely to wrongly flag black defendants as potential repeat offenders compared to the rate for whites. The system also tended to mislabel white defendants as low risk more often than black defendants.

The maker of COMPAS, Northpointe, argued that ProPublica's methodology was wrong. ProPublica has stood by its reporting.

Google ad setting

Google allows users to have a say over what kind of ads they see when they are online. Through the Ad Settings feature, users can include or exclude topics on which they would like to see ads, and can also provide information about themselves so that Google's algorithms can tailor the ads accordingly.

Researchers at Carnegie Mellon in 2015, however, found that a user who specified her gender as female would see fewer instances of an ad from a career-coaching agency promising large salaries than a user whose gender was set as male.

This did not necessarily imply that Google's ad-matching algorithm was discriminatory, because the bias could have come from the ad buyer.

But the researchers argued that it nevertheless showed how personal data provided by a user was being used to discriminate in some manner within the online ad ecosystem.

Google Photos

In 2015, Brooklyn programmer Jacklyn Alciné uploaded a number of photos of himself and a friend, both black, onto Google's Photos app. To his surprise, the app, which relies on a smart algorithm to identify objects in pictures, labelled Alciné and his friend as "gorillas".

"I do have a few questions, like what kind of images and people were used in their initial priming that led to results like these," Mr Alciné told the BBC.

A Google executive was quick to apologise, saying that the bug was "100 percent not OK". While Google was trying to work out a long-term solution, it applied a stop-gap measure of removing the "gorilla" tag from use.

Compiled by BT