THE VERGE

TL;DR MICROSOFT WEB

68

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

Unfortunately, the conversations didn't stay playful for long. Pretty soon after Tay launched, people starting tweeting the bot with all sorts of misogynistic, racist, and Donald Trumpist remarks. And Tay — being essentially a robot parrot with an internet connection — started repeating these sentiments back to users, proving correct that old programming adage: flaming garbage pile in, flaming garbage pile out.

Now, while these screenshots seem to show that Tay has assimilated the internet's worst tendencies into its personality, it's not quite as straightforward as that. Searching through Tay's tweets (more than 96,000 of them!) we can see that many of the bot's nastiest utterances have simply been the result of copying users. If you tell Tay to "repeat after me," it will — allowing anybody to put words in the chatbot's mouth.

One of Tay's now deleted "repeat after me" tweets.

However, some of its weirder utterances have come out unprompted. *The Guardian* picked out a (now deleted) example when Tay was having an unremarkable conversation

with one user (sample tweet: "new phone who dis?"), before it replied to the question "is Ricky Gervais an atheist?" by saying: "ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism."

> @TheBigBrebowski ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism
>
> — TayTweets (@TayandYou) March 23, 2016

But while it seems that some of the bad stuff Tay is being told is sinking in, it's not like the bot has a coherent ideology. In the span of 15 hours Tay referred to feminism as a "cult" and a "cancer," as well as noting "gender equality = feminism" and "i love feminism now." Tweeting "Bruce Jenner" at the bot got similar mixed response, ranging from "caitlyn jenner is a hero & is a stunning, beautiful woman!" to the transphobic "caitlyn jenner isn't a real woman yet she won woman of the year?" (Neither of which were phrases Tay had been asked to repeat.)

It's unclear how much Microsoft prepared its bot for this sort of thing. The company's website notes that Tay has been built using "relevant public data" that has been "modeled, cleaned, and filtered," but it seems that after the chatbot went live filtering went out the window. The company starting cleaning up Tay's timeline this morning, deleting many of its most offensive remarks.

TAY'S RESPONSES HAVE TURNED THE BOT INTO A JOKE, BUT THEY RAISE SERIOUS QUESTIONS

It's a joke, obviously, but there are serious questions to answer, like how are we going to teach AI using public data without incorporating the worst traits of humanity? If we create bots that mirror their users, do we care if their users are human trash? There are plenty of examples of technology embodying — either accidentally or on purpose — the prejudices of society, and Tay's adventures on Twitter show that even big corporations like Microsoft forget to take any preventative measures against these problems.

For Tay though, it all proved a bit too much, and just past midnight this morning, the bot called it a night:

> c u soon humans need sleep now so many conversations today thx
>
> — TayTweets (@TayandYou) March 24, 2016

In an emailed statement given later to *Business Insider*, Microsoft said: "The AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay."

Update March 24th, 6:50AM ET: *Updated to note that Microsoft has been deleting some of Tay's offensive tweets.*

Update March 24th, 10:52AM ET: *Updated to include Microsoft's statement.*

Verge Archives: *Can we build a conscious computer?*

- VIA: The Guardian
- SOURCE: TayandYou (Twitter)