gizmodo.com

No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart

Tom McKay

6-7 minutes

Photo: AP

In recent weeks, a story about experimental Facebook machine learning research has been circulating with increasingly panicky,

Skynet-esque headlines.

“Facebook engineers panic, pull plug on AI after bots develop their own language,” one site wrote. “Facebook shuts down down AI after it invents its own creepy language,” another added. “Did we humans just create Frankenstein?” asked yet another. One British tabloid quoted a robotics professor saying the incident showed “the dangers of deferring to artificial intelligence” and “could be lethal” if similar tech was injected into military robots.

References to the coming robot revolution, killer droids, malicious AIs and human extermination abounded, some more or less serious than others. Continually quoted was this passage, in which two Facebook chat bots had learned to talk to each other in what is admittedly a pretty creepy way.

> Bob: I can i i everything else
>
> Alice: balls have zero to me to me to me to me to me to me to me to me to
>
> Bob: you i everything else
>
> Alice: balls have a ball to me to me to me to me to me to me to me to me

The reality is somewhat more prosaic. A few weeks ago, FastCo Design did report on a Facebook effort to develop a “generative adversarial network” for the purpose of developing negotiation software.

The two bots quoted in the above passage were designed, as explained in a Facebook Artificial Intelligence Research unit blog post in June, for the purpose of showing it is “possible for dialog agents with differing goals (implemented as end-to-end-trained

neural networks) to engage in start-to-finish negotiations with other bots or people while arriving at common decisions or outcomes."

The bots were never doing anything more nefarious than discussing with each other how to split an array of given items (represented in the user interface as innocuous objects like books, hats, and balls) into a mutually agreeable split.

The intent was to develop a chatbot which could learn from human interaction to negotiate deals with an end user so fluently said user would not realize they are talking with a robot, which FAIR said was a success:

> "The performance of FAIR's best negotiation agent, which makes use of reinforcement learning and dialog rollouts, matched that of human negotiators ... demonstrating that FAIR's bots not only can speak English but also think intelligently about what to say."

When Facebook directed two of these semi-intelligent bots to talk to each other, FastCo reported, the programmers realized they had made an error by not incentivizing the chatbots to communicate according to human-comprehensible rules of the English language. In their attempts to learn from each other, the bots thus began chatting back and forth in a derived shorthand—but while it might look creepy, that's all it was.

"Agents will drift off understandable language and invent codewords for themselves," FAIR visiting researcher Dhruv Batra said. "Like if I say 'the' five times, you interpret that to mean I want five copies of this item. This isn't so different from the way communities of humans create shorthands."

Facebook did indeed shut down the conversation, but not because they were panicked they had untethered a potential Skynet. FAIR

researcher Mike Lewis told FastCo they had simply decided “our interest was having bots who could talk to people,” not efficiently to each other, and thus opted to require them to write to each other legibly.

But in a game of content telephone not all that different from what the chat bots were doing, this story evolved from a measured look at the potential short-term implications of machine learning technology to thinly veiled doomsaying.

There are probably good reasons not to let intelligent machines develop their own language which humans would not be able to meaningfully understand—but again, this is a relatively mundane phenomena which arises when you take two machine learning devices and let them learn off each other. It’s worth noting that when the bot’s shorthand is explained, the resulting conversation was both understandable and not nearly as creepy as it seemed before.

As FastCo noted, it’s possible this kind of machine learning could allow smart devices or systems to communicate with each other more efficiently. Those gains might come with some problems —imagine how difficult it might be to debug such a system that goes wrong—but it is quite different from unleashing machine intelligence from human control.

In this case, the only thing the chatbots were capable of doing was coming up with a more efficient way to trade each others’ balls.

Exclusive: *Infinity Train*’s New Comic-Con Trailer Promises One Hell of a Wild Ride

Saturday 1:50PM

There are good uses of machine learning technology, like improved medical diagnostics, and potentially very bad ones, like riot prediction software police could use to justify cracking down on protests. All of them are essentially ways to compile and analyze large amounts of data, and so far the risks mainly have to do with how humans choose to distribute and wield that power.

Hopefully humans will also be smart enough not to plug experimental machine learning programs into something very dangerous, like an army of laser-toting androids or a nuclear reactor. But if someone does and a disaster ensues, it would be the result of human negligence and stupidity, not because the robots had a philosophical revelation about how bad humans are.

At least not yet. Machine learning is nowhere close to true AI, just humanity's initial fumbling with the technology. If anyone should be panicking about this news in 2017, it's professional negotiators, who could find themselves out of a job.

[Fastco Design]