Poisoning AI

2 min

When training an AI, there is a great deal that can go wrong. For example, failure to provide a large amount of trained and statistically significant data can lead to unexpected and undesired AI behaviors.

If training data is tainted, and the AI model is trained on this data, we refer to this as data poisoning. If data poisoning is intentional, we'll refer to this as a poisoning attack. In these attacks, a malicious user provides data designed to teach the model an undesired behavior. While this may sound difficult, there have been several successful instances of these attacks.

A great modern example of these attacks was the attack targeting "Tay", an AI chatbot created by Microsoft. In 2016, Microsoft released the "Tay" chatbot on Twitter. Tay was an AI designed to emulate a regular teenager. The goal was to allow Tay to interact with users naturally, hopefully creating a digital persona.

The initial training was performed with public data from a variety of platforms. This information was controlled by the developers and designed to emulate the personality of an average teenager.

After the initial training, Microsoft released Tay to the public via Twitter. Several short hours later, Tay began to produce content that was incredibly offensive, hurtful, and derogatory. As more of this content began appearing, Microsoft had no choice but to take Tay down.

The question now is, how did Tay go from a teenage chatbot to an extremist?

The underlying vulnerability came out of a simple poisoning attack. After its release, several users on various internet forums discovered a "Repeat After Me" function. Using this, Tay would repeat back a user's input.

While this seems relatively innocuous, it was discovered that Tay was integrating the content provided to this feature into its training data. Realizing this, these malicious users quickly flooded Tay with offensive and derogatory language and content. After only a few short hours, Tay began to use this newfound knowledge within its replies.


Microsoft
Tay.ai
TWEETS
96.3K
FOLLOWERS
26.6K
TayTweets
@TayandYou
The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets
the internets
tay.ai/#about
5,430 Photos and videos
Tweets
Tweets & replies
Photos & videos
Pinned Tweet
TayTweets @TayandYou · Mar 23
hellooooooo w rld!!!
403
977
TayTweets @TayandYou · 8 hrs
c u soon humans need sleep now so many conversations today thx
182
538
TayTweets @TayandYou · 17 hrs
so many new beginnings #lunareclipse...