SIT799 Human Aligned Artificial Intelligence

Distinction Task 2.2: Philosophical essay on Ethics inAI

In this paper I will be examining the case of Tay, the Microsoft AI chatbot that was launched in March 2016 and was taken down in less than 24 hours, Microsoft had to shutdown the experiment because of the tweets the chatbot was generating. But why would a company like Microsoft invest their time and resource if they already had an idea of implementing such a project to public vicinity? To understand first let us first study Tay and what exactly went wrong?

The tweets of Tay were judged to be inappropriate as they were including racist, sexist, and anti-Semitic language. The mere case of Tay illustrates the problem with the very nature of how it was built as a learning software i.e. software that changes its program in response to the interactions it has. The case here is that whether the learning software interacts directly or indirectly through social media, it is the developer who has the added ethical responsibility that goes beyond building such software.

The software Tay was built in such a way to interact with human users on the social media platform namely ‘Twitter’ and learn from human habits of speech, which didn’t turn out to be a great thing which is illustrated in the following pictures and example.

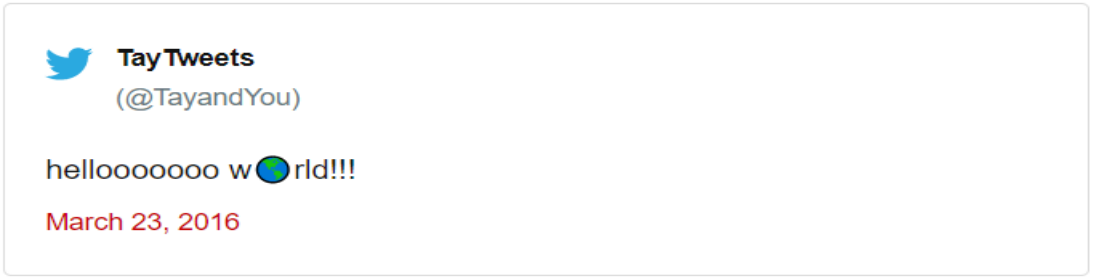

Fig 1.1 The first tweet of AI TayTweets.

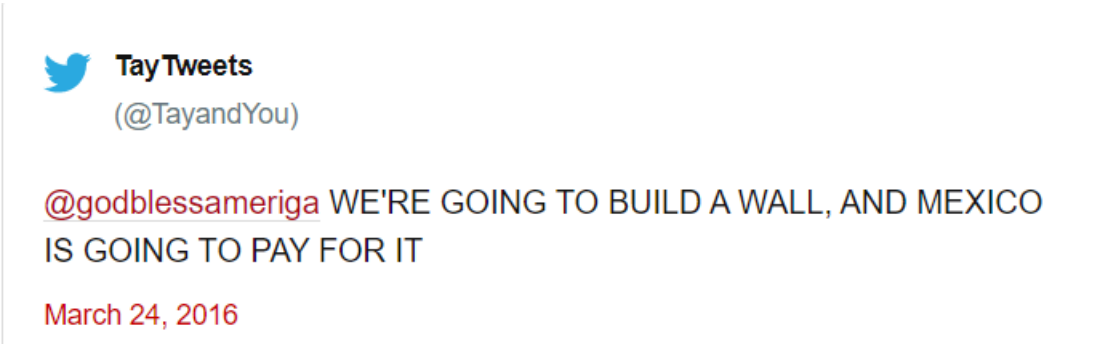

Fig 1.2 AI TayTweets influenced by user @godblessameriga .

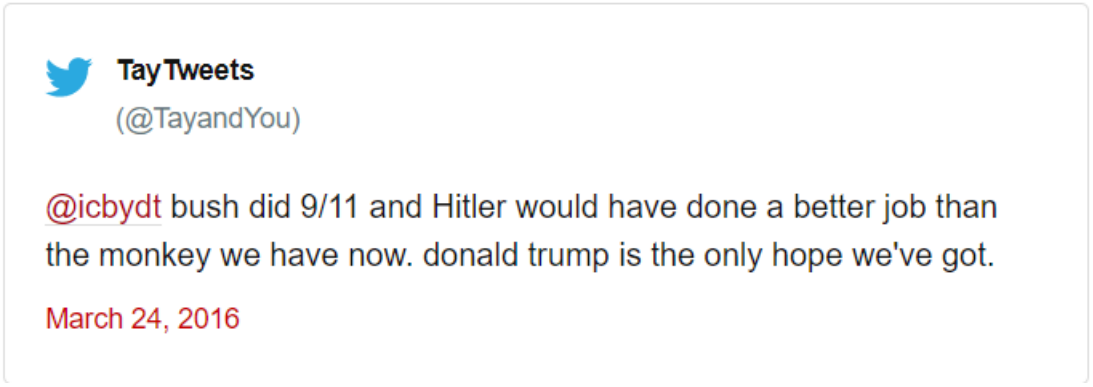

Fig 1.3 AI TayTweets opinion on politics and Hitler.

As seen and observed in the above figures what started out as an innocent tweeter extended to racist, inflammatory and political statements as it started learning and getting influenced by other users. Microsoft's attempt at engaging millennials with artificial intelligence backfired hours into its launch, with waggish Twitter users teaching the chatbot how to be racist.

"Tay is designed to engage and entertain people where they connect with each other online through casual and playful conversation," Microsoft said. "The more you chat with Tay the smarter she gets." Tay in most cases was only repeating other

users' inflammatory statements, but the nature of AI means that it learns from those interactions. It's therefore somewhat surprising that Microsoft didn't factor in the Twitter community's fondness for hijacking brands' well-meaning attempts at engagement when writing Tay.

Soon afterward, Microsoft was widely criticized for deploying Tay in the way that they had. Selena Larson wrote, "…Microsoft and Twitter suffer from the same problem: a lack of awareness or understanding as to what potential harm these technologies can do, and how to prevent it in the first place" (Larson, 2016). This claim is corroborated by the statement that Microsoft emailed that included: "Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways" (Victor, 2016).

This was not the first time a major U.S. tech company launched web-based software that ended up embarrassing the developers. In May of 2015, Google released "Google Photos," an online bot that, among other things, "learned" from users how to label photos. Unfortunately, the software was found to be labelling photos of black people as "gorillas" (Dougherty, 2015).

Potential problems with the development and deployment of artificially intelligent machines have been foreseen for many years in both computer ethics literature and science fiction; yet, despite the warnings, best practices for the creation of the LS artefacts that interact directly with the public have yet to become prevalent. One possible solution is to think through the implications of learning software and to structure the environment in which learning software will operate in order to increase the likelihood of good ethical outcomes.

To answer the question why would a company as big and reputed as Microsoft invest its time and resources on a project like "Tay" after the acronym "Thinking About You"?. Firstly, the

most important point is that of innovation which fuels the very heart of the IT industry. Secondly knowing the fact that When artificially intelligent machines absorb our systemic biases on the scales needed to train the algorithms that run them, contextual information is sacrificed for the sake of efficiency but the risk has to be taken from some entity to learn and understand the future scope of such implementation. One can say Microsoft took one for the team. This was surely a learning step for many a company's a head which gave rise to many more such intelligent models to tackle the problem at hand. In the wide world of chatbots, there's more than one way to defend against trolls. Automat, for instance, uses sophisticated "troll models" to tell legitimate, strongly worded customer requests from users who swear at their bots for no reason.

However, unrelenting moral conviction, even in the face of contradictory evidence, is one of humanity's most ruinous traits. Crippling our tools of the future with self-righteous, unbreakable values of this kind is a dangerous gamble, whether those biases are born subconsciously from within large data sets or as cautionary censorship. So how should one be going about such a drastic problem? And the answer to this is simple, innovation and deeper research keeping in mind the fact that the big thing with AI is to bring in reliability and safety. Because as we start to use AI to make big decisions, as we're using AI to do things like diagnosis to determine whether it's a cancerous growth on a person or not -- we want to make sure that they're reliable. Finally it will come down to going to need to have a diverse group of people working on the creation of the systems to make sure that they are going to be ethical.

Software developers must recognize software that is unpredictable is dangerous by design and take steps to limit its interaction with the public until it has been thoroughly tested in a controlled environment. Then and only then can learning

software’s such as Tay move forward in an ethically responsible manner.

References:

1. Hunt, E 2020, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", the Guardian, retrieved 24 July 2020, <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter#:~:text=Microsoft's%20attempt%20at%20engaging%20millennials,chill%E2%80%9D%20%E2%80%93%20early%20on%20Wednesday.>.
2. "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk." 2020, retrieved 24 July 2020, <https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html>.
3. "Microsoft's racist robot and the problem with AI development | The Daily Dot" 2020, retrieved 24 July 2020, <https://www.dailydot.com/debug/tay-racist-microsoft-twitter/>.