HANZE UNIVERSITY OF APPLIED SCIENCES

ACADEMIC WRITING

INFORMATION COMMUNICATIONS AND TECHNOLOGY

Root cause analysis of Microsoft's Tay AI Bot

Authors
J. MELLEMA
R. VOETMAN

Assecor
Y.TIELEMAN

March 29, 2020

Hanzehogeschool Groningen
University of Applied Sciences

Research Report

Root cause analysis of
Microsoft's Tay AI Bot

Authors	R. Voetman (Student)	J. Mellema (Student)
Student number	388007	386693
University mail address	r.voetman@st.hanze.nl	j.m.mellema@st.hanze.nl
Study	Information Communications and Technology	
Assecor	Y.Tieleman	
Class	ITV2C	
Academic Year	2019 - 2020	
Module	Software Engineering	
Course	Academic Writing	
Publication date	March 29, 2020	
Version	1.11	

REVISION HISTORY

Version	Date	Description	Authors
0.1	8 March 2020	Designed a prototype of the general layout of the document	R. Voetman J. Mellema
1.0	11 March 2020	General layout complies with the standards prescribed by the University	J. Mellema
1.1	14 March 2020	Enhanced layout by converting the document to the LaTeX file format.	R. Voetman
1.2	15 March 2020	Wrote the problem identification	R. Voetman
1.3	17 March 2020	Wrote the Introduction and the preface	J. Mellema
1.4	18 March 2020	Defined the problem statement, objective and research questions.	R. Voetman J. Mellema
1.5	18 March 2020	Defined the concept "Ethical software development" in the theoretical framework.	R. Voetman
1.6	21 March 2020	Conducted and composed the Ethical Analysis	R. Voetman
1.7	21 March 2020	Defined the concept "Weak Artificial Intelligence" in the the-oretical framework.	J. Mellema
1.8	23 March 2020	Conducted the technical analysis.	J. Mellema
1.9	25 March 2020	Concluded the analysis which serves as a basis for the recommendation.	R. Voetman J. Mellema
1.10	26 March 2020	Formatted the bibliography in LaTeX format.	J. Mellema
1.11	27 March 2020	Formatted the Bibliography using APA-style formatting.	R. Voetman J. Mellema

PREFACE

In the post propaedeutic phase of the Bachelor's programme Information Communications and Technology, we conducted a root cause analysis on a failed artificial intelligence or IoT project. This report will provide a root cause analysis on technical aspects as well as ethical aspects of the project. The definitions used in this report have been defined in the Theoretical Framework. The technical analysis and the ethical analysis will be presented in the Conducted Analysis section. Furthermore, the conclusion summarizes the conducted analysis. Finally, the conclusion will serve as a basis for our recommendations.

CONTENTS

1. Introduction

Tay, the artificial intelligence Twitter chatbot which was taken down 16 hours after its launch, was a project being developed by Microsoft Corporation. The chatbot was influenced by the design of Xiaoice, another chatbot developed by Microsoft and deployed in China. In 2015, Xiaoice chatted with 20 million users successfully, without any incidents. However this was not the case for Tay. Tay was taken offline since it started tweeting anti-semitic and racist remarks. Microsoft stated: "we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways."[**?**]

This report will identify and research the technical as well as the ethical factors that could have led to the unexpected behaviour of the chatbot. Which led to it being taken offline. Microsoft Corporation requested a thorough root cause analysis on Tay AI. The conducted research serves as a recommendation for Microsoft Corporation. regarding the development of future artificial intelligence chatbots.

Firstly, the theoretical framework will elaborate on two main topics of this report. Weak artificial intelligence and ethical software development. It will describe the chosen definitions and used sources of this report. Secondly, we will present our analysis in the Conducted Analysis section. This consists of a technical analysis as well as an ethical analysis. Furthermore, the findings will be summarized in the conclusion and it will explain which elements of an artificial intelligence chatbot should be taken into careful consideration when developing and testing an intelligent chatbot.

2. Problem Characterization

In the first hours of the deployment of Tay, the responses of the chatbot were achieving more accuracy as more users interacted with the chatbot. The chatbot would base its answers on the vocabulary used in the conversations it had with its followers. After being online for several hours, the chatbot received more attention which led to an increasing number of twitter followers. This increase provides evidence of the risk of trolls, who could try to manipulate the chatbot.

After 16 hours, Tay was taken offline because it had expressed immoral thoughts which violated the guidelines of twitter. Moreover, it was clear that the behaviour was inappropriate so it had to be taken down, as a lot more people could have been offended/ hurt by those tweets. Microsoft stated that trolls[**?**], caused Tay to tweet those messages[**?**]. Which has supporting evidence, since Tay had to be "exposed" to those remarks by someone or something.

3. Problem identification

The chatbot (Tay) in itself did not malfunction on a technical level. The software did not contain flaws that resulted in a technical failure (e.g. a disability to post new tweets). On the contrary, it was not Microsoft their intention to make the bot post tweets to could be categorized as offensive. The problem therefore has a much larger ethical aspect which has to be considered.

A question that arises is, could Microsoft have prevented this behavior by, for example, applying a filter to Tay's communication layer? Although Microsoft states that they "implemented a lot of filtering and conducted extensive user studies with diverse user groups."[**?**] it remains unknown to what extend these tests were executed. In addition, if the conducted user studies were extensive, the vulnerability could possibly have been discovered at an earlier stage. The possibility arises that the developers knew about the vulnerability prior to the release of Tay.

3.1 Problem statement/ Objective

The **problem statement**: "After 16 hours, Tay was taken offline for expressing anti-Semitic, racist and discriminating thoughts." The **objective** of this rapport: "Identify and research the technical as well as the ethical factors that could have led to the unexpected behaviour of Tay."

3.1.1 Research questions

Main question: How could Tay's algorithms have been abused after the extensive testing of Microsoft corp.?

Sub-questions:

- What factors could possibly have influenced Tay to behave in such an unethical manner?
- Did the developers know about the vulnerability prior to the release of Tay?

4. Theoretical Framework

4.1 Weak Artificial Intelligence

In order to properly understand the basis of Tay, the artificial intelligence chatbot (AI chatbot), it is necessary to differentiate two major concepts in the AI paradigm. Weak and strong AI. As the names suggest the difference lies in the complexity of the algorithm.

A strong AI is able to think independently, and make decisions that are not predefined by the creator. A strong AI can be categorized as "systems that exhibit autonomous intelligence and decision making"[**?**].

A weak AI is built with the idea of replication/ duplication in mind. The concept of a weak AI is the opposite of the concept of a strong AI. Weak AI is used for automatable tasks or tasks that can be repeated with easy. It is meant to perform a specific task without any independent thinking.

Therefore, a weak AI can excel in a task and master it within a shorter time span than any human could. From this point of view, a weak AI could be viewed as an "intelligent" algorithm. However this is usually not true, since it can only perform a task well and it is not able to think independently and it is not able to "learn" a skill by itself.

There is no general consensus on the exact definition of weak AI. However, a definition that most researchers acquiesce upon is: "Weak AI is the concept that whatever the program is meant to do, it is merely trying to replicate or duplicate that function, and for most tasks that is sufficient"[**?**]. Since a rough definition of weak AI is sufficient in order to understand the working of the chatbot, the definition of Searle will be used in this report.

4.2 Ethical software development

Over the years, multiple guidelines have been drawn up for Code Of Ethics by industry-trusted institutions and organizations. A subselection of this wide array of documents it listed below:

- IEEE (Institute of Electrical and Electronics Engineers) Code of Ethics[**?**]
- ACM (Association for Computing Machinery) Code of Ethics and Professional Conduct[**?**]
- CEI (Computer Ethics Institute) Ten Commandments of Computer Ethics[**?**]

Although these are professional guidelines that are respected by the industry, they do not provide a succinct definition of the term "Computer Ethics". A succinct but well-defined definition of "Computer Ethics" is given by a Professor of Philosophy at Dart-Mouth College: “Computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology.”[**?**]. The definition of Moore will be used in this report, mainly because it speaks of the "social impact of computer technology". This fits in well with the problem statement given that Tay was released on a social media platform (i.e. Twitter).

5. Conducted Analysis

5.1 Technical analysis

The technical analysis consist of an intelligence analysis of Tay. This report cannot thoroughly conduct a technical analysis on the implementation of the artificial intelligence chatbot, since the project is closed sourced. Meaning that the code and implementation is not available for the general public. Therefore, alternative sources had to be consulted. The intelligence analysis was conducted by students of the Manipal Institute of Technology in India. This research provides examples of the how the chatbot "learned" from the input it received.

The conducted intelligence analysis consists of analyzing its 3000 most recently tweeted tweets. The findings which will be elaborated on are: Taxonomy Classification, Frequency Analysis and Term Frequency-Inverse Document Frequency vectors. We will provide further explanation in order to make the research comprehensible.

Firstly, the chatbot its tweets were categorized into predefined categories. "The TayTweets corpus is analysed as a whole and five different categories are inferred, with corresponding confidence measures for each classification."[**?**]. These results led to an overview which primarily consisted of everyday topics. Which provides evidence of the audience of Tay, regular Twitter users.

Secondly, a frequency analysis was conducted on the tweets. Its vocabulary mainly consisted of basic phrases and slang used on Twitter. Some examples are: "chatting", "human", "DM", "keep"[**?**]. This analysis provides evidence of implementation of Tay. It supports the fact that the responses Tay generated were based on the questions it was asked.

Lastly, term frequency-inverse document frequency vectors were determined in order to research whether there exists a similarity between its responses and the questions it was asked. "We convert both of these corpora into a Term Frequency-Inverse Document Frequency vectors and then compute the cosine similarity."[**?**]. A cosine similarity of 0.9640545176 out of 1.0 was obtained. "A score closer to 1 indicates that both the questions and answers were highly similar"[**?**]. This experiment provides evidence its responses and the questions Tay was asked.

In conclusion, the conducted intelligence analysis supports the fact that Tay uses the input of the user in order to formulate a fitting response.

5.2 Ethical analysis

A question that arose in the problem identification section is, did the developers of Tay know about the vulnerability prior to the release of Tay. The ACM Code of Ethics says that a developer should "give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks" [**?**]. This means that if they had not foreseen this endangerment, they would still be obligated to foresee that Tay could eventually behave in a way they did not anticipate.

It cannot be guaranteed for certainty that all the developers were aware of these possible consequences. However, it can be concluded that Microsoft Corporation as a company was aware of the possible risks. The statement that Peter Lee gave on behalf of the company states that they "implemented a lot of filtering and conducted extensive user studies with diverse user groups."[**?**].

When examining the results an argument can be made whether filtering was implemented. Filtering could not have prevent this situation completely, but by, for example, blocking anti-Semitic language the consequences of this exploit could have been prevented to a large extend.

To summarize, an argument could be made about on which ethical theory Microsoft Corporation based their acting. It is safe to say that Microsoft knew of the possible consequences and that apparently a decision has been made to release Tay. This is a common example of utilitarianism because the potential risk of this vulnerability being exploited was outweighed by the ability to collect a lot of real life data to improve their own algorithms.

6. Conclusions

In conclusion, the technical analysis supports the argument that Tay uses the input of its users in order to formulate a fitting response. As identified by the intelligence analysis, Tay learns from its users and there does not seem to be any limit in what could have been taught to Tay. Furthermore, the ethical analysis shows that developers of self learning software should give comprehensive and thorough attention to the potential unforeseen vulnerabilities because it could behave in a way that was not anticipated.

We would recommend implementing extensive filtering, common taboo and racist remarks are some examples. We think that a general rule of thumb would be words that are known to be discriminating or racist. In addition to extensive filtering, we would also suggest in developing a test framework where all kinds of topics and remarks will be discussed with the chatbot. This is in order to foresee that self learning software could behave in a way which was not anticipated at first.

APPENDIX A

Peer Review - Roy Voetman, Jordi Mellema - ITV2C - Academic Writing

Jordi peer reviewed the following elements of the report:

- 3. Problem Identification
- 4.2 Theoretical Framework - Ethical software development
- 5.2 Conducted Analysis - Ethical Analysis

These chapters were written by Roy. I have summarized my findings below:

3. Problem Identification:

'e.g. a disability ability to post new tweets' not entirely sure what the meaning of this sentence is.

"It was not Microsoft**'s** intention". I believe it is stated in the academic writing handout, that contractions should be avoided when writing an academic report. Therefore I would suggest changing it to: "It was not Microsoft their intention".

"if the conducted user studies really **where** extensive the vulnerability could possibly have been discovered at an earlier stage." The word 'where' should be replaced with "were", furthermore I would advise adding a comma between "extensive" and "the vulnerability". Finally, I would remove "really" since it does not add any more meaning to the sentence.

"Did the developers **knew** about the vulnerability prior to the release of Tay?" 'Did ... knew' should be 'Did ... know'.

How could Tay's algorithms have been abused after the extensive testing of Microsoft **corp.**? I would change corp. to Corporation. As stated on the website of the University of New England: "Avoid using common abbreviations. It is best to write the full term in the text of your writing.". Source: https://aso-resources.une.edu.au/academic-writing/usage/shortened-form-of-words/

4.2 Theoretical Framework: Ethical software development:

I did not find any mistakes nor any erroneous grammatical structures in this subsection. I think this is a well-written subsection, in my opinion. Several guidelines regarding the development of Computer Ethics are listed. Which is more general. Later on, Roy elaborates on the reason why the definition of professor J. H. Moor is sufficient enough. Which is in this is a specific example.The structure from general to specific fits the purpose of this part of the theoretical framework in my opinion. In conclusion, I think this is a well-written subsection.

5.2 Conducted Analysis: Ethical Analysis:

"However, it can be concluded that Microsoft **inc.**" and "ethical theory Microsoft **inc.**" As stated earlier: "Avoid using common abbreviations. It is best to write the full term in the text of your writing." This rule of thumb also applies here. However, I think it would be better to use Microsoft Corporation since using two different terms could be confusing. Furthermore it is also inconsistent.

"This **mean** that if they had not foreseen this **endanger**, they would still be obligated to foresee that Tay, because she learns from her own environment, could eventually behaved in a way they did not anticipate."

Firstly, I think this sentence is quite lengthy. I think you could divide it into several sentences. Secondly, 'mean' should be 'means'. Thirdly, I do not think 'endanger' is appropriate here, I would replace it with 'endangerment'.

Furthermore, "could eventually behaved in a way", seems to be incorrect. I think it should be: "could eventually behave in a way...".

"However, it can be concluded that Microsoft inc. as a company **was** aware of the possible risks **was**." Was is used twice here. The last one should be removed.

"When examining the results it can be concluded that filtering was either not implemented or to a very low extent." I believe this is a logical fallacy. A hasty generalization in this case. The source code of Tay is closed sourced, meaning that the implementation of Tay is not available for the general public.

Making the statement that filtering was not implemented or to a **very** low extent. Is a hasty generalization. Because the implementation of Tay is not available to us. Therefore, a test framework could have been developed that excluded anti-semitic remarks. And included some other slurs or racial remarks. You can't make such a statement because you know too little of the actual implementation, in my opinion.

I would also remove very, as it does not add any meaning (useful) to the sentence.

Roy peer reviewed the following elements of the report:

- 1. Introduction
- 2. Problem Characterization
- 4.1 Theoretical Framework - Weak artificial intelligence
- 5.1 Conducted Analysis - Technical analysis

These chapters were written by Jordi. I have summarized my findings below:

1. Introduction:

"another chatbot developed by Microsoft which is based in China." when reading this sentence it seems like it Microsoft is based in China.

"Tay was taken offline since it started tweeting anti-semitic and racist remarks." Before this sentence it is never mentioned that Tay is a twitter bot.

"Microsoft Corporation. requested for a thorough root cause analysis on Tay AI" Corporation does not need a dot at the end since it is not an abbreviation. "requested for a" should be "requested a".

"Finally, the recommendation will explain which elements of an artificial intelligence chatbot should be taken into careful consideration when developing and testing an intelligent chatbot." Conclusion and recommendation are now merged and not two separate sections.

2. Problem Characterization:

"This increase provides evidence of the risk of trolls, who could try to manipulate the chatbot." Trolls is a very informal word, instead you could describe what a trolling user is.

"Moreover, It was clear that " It without a capital I.

"Microsoft stated that trolls caused Tay to tweet those **thoughts**." Is thought the right word to be used here? To prevent an argument that states if Tay can have thoughts or not it can be changed to the verb "messages" Secondly, there is no citation to this statement.

4.1 Theoretical Framework - Weak artificial intelligence

"Strong AI are able to think independently ..." Strong AI is singular but "are" is used which suggests using the plural form.

"The concept of strong AI is therefore quite the opposite of that of weak AI." Weak AI has not yet been defined, so this comparison can be interpreted as unclear

“Weak AI are built with ...“ Weak AI is singular but “are” is used which suggests using the plural form.

5.1 Conducted Analysis - Technical analysis

“The conducted intelligence analysis consists of analyzing **the its 3000** most recently tweeted tweets.” The bold part is grammatically incorrect, I would remove the “its”.

“Firstly, **its** tweets were categorized into predefined categories.” This is the first sentence and “its” should refer to something in a past sentence, since there isn’t any this sentence is grammatically incorrect.

“Which provides evidence of audience **the Tay**. Everyday Twitter users.” Not really sure what the sentence should say but “the Tay” is not correct.

“Secondly, a **Frequency Analysis** ...” no need to use capitals in the bolded part.