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 15, 2020

Hanzehogeschool Groningen
University of Applied Sciences

**Research Raport**

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 15, 2020 | |
| Version | 0.1.1 | |

# 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 | 10 March 2020 | Wrote the basic structure for the preface. | J. Mellema |
| 1.1 | 11 March 2020 | General layout complies with the standards prescribed by the University | J. Mellema |

# PREFACE

In the post propaedeutic phase of the Bachelor's programme Information Communications and Technology, we had to conduct a root cause analysis on a failed AI or IoT project. This report will provide an analysis on technical aspects as well as ethical aspects of the project, the analysis will be supported by a theoretical framework. The conclusion summarizes the conducted analysis. Furthermore, it will serve as a basis for our recommendations.

# CONTENTS

# 1. Introduction/ Case Description

-General introduction: introduce the topic, pique the interest of the reader, describe/ explain why the research is relevant.

# 2. Problem Characterization

Microsoft revealed the creation of a new chatbot called 'Tay', on March 23 2016. Tay was named after the acronym: "Thinking about you". The chatbot interacted with twitter users and learned from the users' input, it did so using frequency analysis. Frequency analysis consists of evaluating phrases and words, more frequently appearing phrases would have a heavier weight. And thus would be used more by the bot when engaging in conversations.

After 24 hours, however, the bot was taken offline for expressing anti-Semitic, racist and discriminating thoughts. For example, when asked if Tay supported genocide, Tay responded with "i do indeed". 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. Trolls and racist twitter users had influenced/ taught Tay to use inappropriate language. This is an example of the abuse of the frequency analysis algorithm.

# 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 ability to post new tweets). On the contrary, it was not Microsoft's 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 says that they "implemented a lot of filtering and conducted extensive user studies with diverse user groups." (Peter Lee, 2016) it remains unknown to what extend these tests were executed. In addition, if the conducted user studies really where extensive the vulnerability could maybe have been discovered at an earlier stage. The possibility arises that the developers knew about the vulnerability prior to the release of Tay.

# 4. Theoretical Framework

## 4.1 Theoretical background

## 4.2 Literature review

# 5. Methodology

## 5.1 Facts and findings

## 5.2 Technical analysis

## 5.3 Ethical analysis

# 6. Conclusions

## 6.1 Summary of result/findings

## 6.2 Hypothesis

# 7. RECOMMENDATIONS

# 8. BIBLIOGRAPHY