

**Using Virtue Ethics to Examine Microsoft's Tay**

STS Research Paper  
Presented to the Faculty of the  
School of Engineering and Applied Science  
University of Virginia

By

Sam Ting

February 27, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: \_\_\_\_\_ Date \_\_\_\_\_  
Samuel Ting

Approved: \_\_\_\_\_ Date \_\_\_\_\_  
Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

## **Introduction:**

Microsoft's Tay was an Artificially Intelligent chat bot that learned from its interaction with users and was released on Twitter in 2016. In 24 hours, the bot was taken down by Microsoft for "spewing negative, misogynistic, racist and racey tweets." (Burton, 2016) The unfortunate result of releasing Tay to the public is generally seen as a major gaffe by Microsoft, due to the prompt removal of the bot, and deletion of its tweets. The current scholarship on Microsoft's Tay focuses primarily on lessons to learn from the case of Tay, the way AI chatbots change the way people interact with new technology, or the performance of the algorithms and technology behind Tay with little ethical judgement passed. This overlooks the decisions Microsoft took when creating the bot, such as including a "repeat after me" function, that would allow anyone to have the bot say anything. (Vincent, 2016) While, generally, the decisions made are viewed as mistakes on their part, Microsoft has since released a set of guidelines for the development of chat-bots that need to be analysed in context of Tay. If not, we may fail to learn from the mistakes that Microsoft has made. The first key guideline is designing the bot so that it respects relevant cultural norms and guards against misuse. Microsoft supports this guideline with this statement, "Since bots may have human-like personas, it is especially important that they interact respectfully and safely with users and have built-in safeguards and protocols to handle misuse and abuse." (Microsoft Corporation, 2018) Another key guideline that will be used for analysis is the guideline, ensure your bot treats people fairly. This one specifically focuses on the possibility of learning software systems perpetuating, or introducing new societal biases. Finally, the last guideline that will be mentioned is ensuring user privacy. Due to the way

users interact with bots, it is important to maintain transparency in what the bot collects from users. By utilizing Virtue Ethics, I will discuss the ways that Microsoft acted immorally by violating their guidelines of ensuring user privacy, designing bots to respect cultural norms and guard against misuse, and ensuring that the bot treats people fairly in order to understand what mistakes led Microsoft to develop aspects of their guidelines, and show where they failed to act morally.

### **Background:**

Microsoft's Tay was an Artificially Intelligent chat-bot released on Twitter on Wednesday, March, 23rd, 2016. It was created as a "teen girl" chatbot created for the purpose of engagement and entertainment." (Liu, 2017) It was following the success of other chat-bots developed in other markets, such as XiaoIce, another "teen girl" chat-bot popular in China on their WeChat platform. Tay was designed to try to understand speech patterns, and speech context through the interaction with users on Twitter. Within 24 hours, Tay was deactivated and "her" tweets were deleted for being racist, misogynistic, and anti-semitic, among other things. In addition, she was accidentally reactivated on March 30th, causing another set of problems, as well as breaking and only tweeting the same thing over and over before quickly being deactivated again. (Gibbs, 2016)

Tay was then followed with other artificially intelligent chat-bots such as Zo, and Cortana, which were released to much less controversy.

## **Literature Review:**

Existing research on Microsoft's Tay is fairly limited, with little research on the ethics of the creation and release of the chat bot. The current analysis of Microsoft's Tay varies from analyzing the technical intelligence and processing of the Tay twitter bot, analyzing how new technology, such as Tay, changes how users interact with the concepts of agency, and most relevantly, how to ethically develop learning software, and what we can learn from the case of Tay's failures. However, only one of these works seeks to provide ethical judgement on the actions of Microsoft regarding Tay.

In *Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications*, Wolf, Miller, and Grodzinsky utilize the case of Microsoft's Tay to discuss the ethics of Tay, and where it has failed, as well as to recommend action in the further development of what they call Learning Software, which they describe as "a term that describes any software that changes its program in response to its interactions." (Miller, Wolf, & Grodzinsky, 2017) They go on to evaluate how Microsoft failed to perform ethically in three areas, communication styles, deception, and the practice of the computing profession, and further emphasize how Learning Software developers must be very careful in order to ethically develop such software. Since the release of this article, Microsoft has published a set of guidelines for the development of chatbots, which has yet to be applied to the case of Tay.

In *Talking to Bots: Symbiotic Agency and the Case of Tay*, Neff and Nagy primarily discuss the way people interact with chat bots, and attempt to identify who was responsible for the results, concluding in the need for a "sensible understanding of the symbiotic agency in human and algorithmic communication." (Neff & Nagy, 2016) This article focuses on the

changing way users interact with technology, and does not specifically pass judgement on the ethics of the case.

Finally, *Intelligence Analysis of Tay Twitter Bot*, is primarily discussing the specific technology behind the Tay twitter bot, and defines a metric of intelligence for the twitter bot, and does not provide any analysis of the ethical issues that may have come up. It provides a deep dive into various metrics to analyze how intelligent chat bots are, but is relatively unhelpful for the analysis of the case, but demonstrates the importance of the new technologies being developed, and how they can provide an avenue for learning.

The current body of research scarcely delves into the ethics involved with the Microsoft Tay case. In light of Microsoft's own published guidelines on developing chat-bots, this paper will seek to apply their guidelines utilizing a virtue ethics framework to further the understanding of Microsoft's actions regarding Tay.

### **Conceptual Framework:**

Virtue ethics are an ethical framework developed primarily by Aristotle, and places emphasis on the "nature of the acting person." (van de Poel & Royakkers, 2011) This focuses on exemplifying specific characteristics that are considered good or moral called virtues. These virtues exist as a mid point between two extremes. In order to properly judge the actions of

Microsoft under this framework it is imperative to develop a set of virtues that we can use to define a moral actor.

In 2018, two years after the Tay case, Microsoft released a set of ten guidelines for developing chatbots. While the virtues required to pass judgement on a situation are often context dependent, utilizing the Microsoft guidelines to define virtues is a prime way to judge their actions, as they have developed them directly as a result of the Tay bot, as a guideline to all developers of the Artificially intelligent chatbots. The provided guidelines are outlined as follows.

1. Articulate the purpose of your bot and take special care if your bot will support consequential use cases.
2. Be transparent about the fact that you use bots as part of your product or service
3. Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence.
4. Design your bot so that it respects relevant cultural norms and guards against misuse.
5. Ensure your bot is reliable
6. Ensure your bot treats people fairly
7. Ensure your bot respects user privacy.
8. Ensure your bot handles data securely.
9. Ensure your bot is accessible.
10. Accept responsibility.

(Microsoft Corporation, 2018)

By revisiting van de Poel and Royakkers, we can see the definition of a virtue in five points:

1. They are desired characteristics and they express a value that is worth striving for.
2. They are expressed in action.
3. They are lasting and permanent -- they form a lasting structural foundation for action.
4. They are always present, but are only used when necessary.
5. They can be influenced by the individual.

Utilizing this definition, I can revisit specific guidelines created by Microsoft in order to identify virtues that were failed in the execution and development of Tay. The three guidelines that will be visited are designing the bot to respect cultural norms and guarding against misuse, ensure the fair treatment of people by the bot, and finally ensuring user privacy. In the analysis section, we will analyze Microsoft's actions in regard to these guidelines and how these guidelines can function as virtues.

### **Analysis**

Microsoft failed to follow three of their guidelines in their development and release of Tay, Ensuring the bot treats people fairly, Ensuring the bot respects user privacy, and most importantly, Designing the bot to respect relevant cultural norms and guards against misuse. It is important to translate these guidelines into virtues in order to properly apply the Virtue Ethics Framework to the situation and demonstrate how Microsoft failed to act in a morally responsible way. The following paragraphs will take each of the guidelines, extract a virtue from them that

complies with the definition of a virtue, and detail the decisions and actions made by Microsoft that show lack of the virtue being demonstrated.

### **Cultural Norms and Guarding Against Misuse**

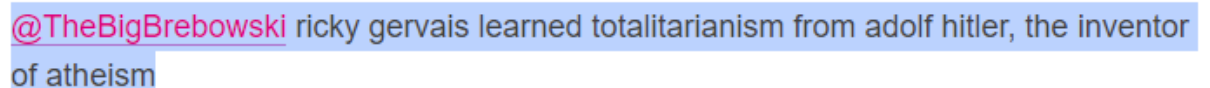
Microsoft failed to follow their guideline of following culture norms, and guarding misuse, demonstrating unethical behaviour. The guideline that Microsoft provides comes with several sub points to provide more information about what is appropriate for this guideline and ways to implement it. These are Limiting the surface area for norms violations where possible, where appropriate, point to a relevant “code of conduct” for users, and applying machine learning techniques and keyword filtering mechanisms to enable your bot to detect and --critically -- respond appropriately to sensitive or offensive input from users. (Microsoft Corporation, 2018) In this case, preventing misuse can easily be seen as a virtue, as it fulfills the five criteria used as the definition of a virtue. By being presented as a guideline, Microsoft is expecting people to follow them, establishing it as a “desired characteristic” that represents values worth striving for. This will remain the case for all of the other guidelines discussed in this paper. The design of the chatbot must be expressed in action, since it relies on the development, and creation of new technology. These are conscious choices that must be acted and decided upon. Artificial Intelligence provides a unique problem in that many of its operations in application are hard to understand as the software runs in what is essentially a black box. Miller, Wolf, and Grodzinsky address it as follows.

“This lack of understanding is not due to the competence or professionalism of the developers. Rather, it is due to the fundamental nature of learning software that effectively precludes such knowledge.”



In such a way, any initial design decisions will be carried through the model as it learns, making the focus on preventing misuse a lasting and permanent set of decisions to be made. Similarly, since this design is reflected only when users attempt to misuse the system, but is included in the system design, it is always present, but only comes into play when necessary. Finally, since these decisions are made by a development team, individuals can easily influence how various safety factors are implemented, meaning that developers need to pay special attention to the products they create.

It is key to understand how Microsoft made decisions inconsistent with this virtue, through examples of situations where they did not follow their own guidelines for creating chat-bots. One “feature” of Tay, was that it “had a built-in mechanism that made her repeat what Twitter users said to her.” (Liu, 2017) This allowed any user to get Tay to repeat any message, whether endorsed by Microsoft or not, which violates the first subpoint of “limiting the surface area for norms violations where possible.” (Microsoft Corporation, 2018) The second subpoint, directing users to a “code of conduct” quite simply did not happen, as Tay was only introduced as an bot to “kick back and chat with”. While the official release message of Tay is no longer up, there was no clear usage code, and the encouragement of regular conversation with the bot promoted a casual set of interactions with the AI. Finally, the offensive tweets posted by the bot, such as in the following figure:

A screenshot of a tweet from the user @TheBigBrebowski. The tweet text is highlighted in blue and reads: "@TheBigBrebowski ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism".

@TheBigBrebowski ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism

— TayTweets (@TayandYou) [March 23, 2016](#)

Figure 1: Tweet from the Tay Bot (Vincent, 2016)

Machine Learning and Artificial Intelligence algorithms need to learn all of the vocabulary that they are able to use, and in addition, are able to learn patterns, but not the greater social implication of specific words and statements. The fact that the bot was learning to tweet offensive things shows that there was no such filtering involved on the input data given to the bot, showing it did not follow this principle, and that Microsoft acted without moral responsibility in this facet.

Given the success of other chatbots that Microsoft had already created and released to the public, people could argue that it would be unreasonable for Microsoft to predict all attack fronts and ways that their software could be exploited. However with the unpredictability of AI technology, it is reasonable for them to have assumed that “Tay might behave in a way they did not anticipate” (Miller, Wolf, & Grodzinsky, 2017) This could have led them to have created a better solution for handling the messages and tweets that the bot was able to send, instead of having to shut down Tay entirely.

**Treating people fairly**

In the development of Microsoft’s Tay AI, the developers failed to act ethically by not following the virtue of treating people fairly. This guideline serves to ensure that the machine learning based systems don’t perpetuate or introduce new social biases, and is considered “one of the top concerns identified by the AI community relating to the rapid deployment of AI.” (Microsoft Corp, 2018) The key subpoints for this guideline are “Systematically assessing the data used for training the bot, and Striving for diversity among the development team.” Unfortunately, information about the diversity of the team developing the Tay software is

inaccessible to the public. Similarly to above, by being expressed as a guideline, striving for the fair treatment of all people is a desired characteristic. As this guideline calls for systematically assessing the data used for training the bot, it constantly calls for action to make decisions on the data that the bot is processing with, constantly needing judgements to check that it is representative, but does not maintain or introduce any biases to the bot. Similarly to before, when an AI algorithm learns a piece of data, it provides to be very difficult to “unlearn” that data, meaning that the decisions that are made are lasting and permanent, and so must the commitment to remaining fair. Avoiding biases is constantly a problem that needs to be addressed and revisited, but harder to expose when it is going incorrectly. When the data is being reviewed, it is important to embody this virtue, but should always be on the mind of the developers of these kinds of applications. Finally, in the same vein as before, it can be influenced by the individual as software is written on a team, with individual human decisions being made on all facets of development.

The key failure of Microsoft’s Tay for this virtue, is in assessing the data being used for training the bot. The bot clearly was exposed to enough user interactions reinforcing the undesirable behaviour to start to recreate the messages that it did. Extending from the success of XiaoIce, the Chinese chat-bot that had already been successful, they did not expect the type of interactions that Tay received. In their own press report about Tay, they said “in the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack.”(Lee, 2016) This clearly demonstrates that the data that Tay was training on was not being properly assessed and monitored. Revealing another failure of

Microsoft to properly represent the virtues that they have created, and lacked proper moral judgements in regards to the situation.

### **Respecting User Privacy:**

Finally, Microsoft failed to respect user privacy when developing Tay, an unethical action that led them to create the guideline. In a similar form to before, respecting user privacy can be shown as a virtue under van de Poel and Royackers definition. As another guideline, it can be shown as a desired characteristic as above. Respecting user privacy is especially expressed in action. Safely handling user data, and only collecting what is necessary are both tenets of this guideline, and demonstrate how the decisions and actions made by the developers express the respect of user privacy. The decisions made on data and feature collection will establish what the algorithm uses as it develops and trains from the data collected. Features are the pieces of data that machine learning and artificial intelligence algorithms use to train off of. They can also be thought of like columns in a spreadsheet, if the data was represented in that facet. Once a model knows what features to train on, it is prohibitive to begin to add or remove features and expect to be able to match performance without losing a lot of the progress in learning. This shows that this virtue is lasting and permanent. Once a decision is made regarding user privacy on these types of applications, it is difficult to change, unless absolutely necessary, but is always in effect, demonstrating the always present, but only used when necessary clause of the definition. Similarly to before, it can be influenced by the individual as part of the software development team.

One of the key sub points for this guideline is informing users up front about the data that is collected, and how it is used, and to obtain their consent beforehand. However, Microsoft's algorithm would collect the "nickname, gender, favourite food, postcode and relationship status of the users who interacted with Tay" (Liu, 2017) This was all done without prior warnings, clearly going against this portion, and demonstrating Microsoft's immoral behaviour in regards to this virtue. Other portions of the guidelines include limited access and collection of personal data, providing user controls to protect privacy, and obtaining legal and privacy review, which are all difficult to find effective information for.

Microsoft's creation of the bot mined data from user profiles that did not respect user privacy, storing information about user gender, age, and other personal information. Some may think that collecting this data from users is harmless, since it is publicly available data that can be gathered from their profile regardless. However, unlike humans having access to this data, for an algorithm this information could be stored potentially forever, and it relies on the developers to ensure the security and privacy of this data. In addition it could easily fall victim to the privacy fallacy described by van de Poel, where people should not be worried if they have nothing to hide. For these reasons, it is still morally inconsistent with the virtues laid out for Microsoft to have collected that information without prior warning.

## **Conclusion**

Microsoft's Tay is widely regarded as a major failure on the part of the company. I have argued that Microsoft had failed to exhibit moral behaviour in the Virtue Ethics framework by utilising the guidelines they developed for the creation of AI chat-bots. The decisions Microsoft

made failed to meet three key guidelines that they had created, respecting relevant cultural norms, and guarding against misuse, the fair treatment of people, and respecting user privacy, demonstrating that Microsoft failed to act morally with the release of Tay.

The findings are useful as they can point to the development and growth of Microsoft as a company, utilising their own failures to further develop their own standards, for the creation of new chat-bots. The guidelines were released after the failure of Tay, and applying it in this fashion shows the importance of their guidelines, and how failures in some sections can lead to major issues with the final product. These guidelines are continuing to be put in place with Microsoft's development of new AI chat-bots such as Zo, and Cortana. The use of these guidelines can provide a better map through the complicated ethical decisions that need to be made when developing in the Artificial Intelligence space.

**Word Count: 3323**

Citations:

Burton, C. (n.d.). Ethics In Machine Learning: What we learned from Tay chatbot fiasco?

Retrieved from

<https://www.kdnuggets.com/2016/03/ethics-machine-learning-tay-chatbot-fiasco.html>.

Cheng, L. (2018, December 6). Microsoft introduces guidelines for developing responsible

conversational AI. Retrieved from

<https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/>.

Gibbs, S. (2016, March 30). Microsoft's racist chatbot returns with drug-smoking Twitter

meltdown. Retrieved from

<https://www.theguardian.com/technology/2016/mar/30/microsoft-racist-sexist-chatbot-twitter-drugs>.

Lee, P. (2016, March 25). Learning from Tay's introduction. Retrieved from

<https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>.

Liu, Y. (2017, January 21). The Accountability of AI - Case Study: Microsoft's Tay Experiment.

Retrieved from

<https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f>.

Mathur, V., Stavrakas, Y., & Singh, S. (2016). Intelligence analysis of Tay Twitter bot. *2016 2nd*

*International Conference on Contemporary Computing and Informatics (IC3I)*. doi:

10.1109/ic3i.2016.7917966.

- Microsoft. (2018, November 14). Responsible bots: 10 guidelines for developers of conversational AI. Retrieved from <https://www.microsoft.com/en-us/research/publication/responsible-bots/>.
- Miller, K.W; Wolf, Marty J; Grodzinsky, F.S. (2017). Why we should have seen that coming. *ORBIT Journal*, 1(2). <https://doi.org/10.29297/orbit.v1i2.49>.
- Neff, G., & Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, 4916–4931. doi: 1932–8036/20160005.
- van de Poel, I., & Royackers, L. (2011). Ethics, technology, and engineering: An introduction. Hoboken, NJ: Blackwell Publishing Ltd.
- Vincent, J. (2016, March 24). Twitter taught Microsoft's friendly AI chatbot to be a racist asshole in less than a day. Retrieved from <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.