****

**Student No: 13006860**

**Do as I Say, Not as I as I Do: How Do We Avoid Reproducing Bias in AI and Machine Learning?**

**Module Code: UZRSSR-15-3**

**Date: 30/04/2017**

**Department of Engineering, Design, and Mathematics**

**University of the West of England**

**Contents**

[Introduction 3](#_Toc480973351)

[Machine Learning 5](#_Toc480973352)

[Learning from what? 6](#_Toc480973353)

[Conclusion 9](#_Toc480973354)

[References 10](#_Toc480973355)

# **Introduction**

On the 23rd of March 2016, the chatbot Tay was introduced on to the social media platform, Twitter. Tay which is an acronym for “thinking about you” was developed by Microsoft’s technology and research division. It was designed to mimic the language patterns of an American teenage girl by learning from interactions with other Twitter users. Tay “mined” public data with the use of machine learning algorithms that attempt to further experiment and tackle the issue of conversational understanding in AI. The experiment was a success; Tay began interacting with other users and could formulate responses and reply in contextually relevant ways. Within 16 hours of her release and after 96,000 tweets, Tay was suspended by the Microsoft team. A critical flaw had been exposed, the team had not put any filters on what Tay could learn and consequently reproduce so she began tweeting holocaust denying, racist and conspiratorial messages [1].



Figure : Tay Tweets [1]

Researchers at Carnegie Mellon University have found men were being shown adverts for high paying jobs at a significantly higher rate than women. The team created 1,000 simulated user profiles on Google. Half were designated to be female while the other half male. These simulated users then visited employment sites using Google. Google’s algorithms are designed to infer things about a user based on the sites they visit; this then allows them to target advertising and tailor it specifically to your user profile. The researchers discovered that their male simulated users were shown adverts for high paying jobs 1800 times while the female users were shown adverts 300 times [2]. They found no evidence that Google was purposefully behaving in a discriminatory manner; there is no way for them to determine the cause of the disparities observed as the process of serving ads to users involves complex relationship between Google, the advertisers and the algorithms being used.

The third and final example that serves to illustrate the crux of this report is one that involved Google’s facial recognition technology: a feature which classifies user’s photos into distinct categories for ease of use and access. The algorithm sorted user Jacky Alcine’s photos of himself with friends under the category of gorillas. Referring to people of African descent as gorillas or monkeys is a deeply racist propagandizing trope used to justify their enslavement and degradation. It is certainly not something an algorithm in the 21st century should be doing, yet it happened. Some have theorized that the training data used for the recognition algorithm focused too much on white faces which meant it built a model of the world that would mainly recognise white faces as human [3].

All three examples address the ethical discussion that will be expanded upon in this report. That is, how do we prevent machines and algorithms from reproducing the bias and inequalities we see in the world today? Algorithms that are deployed to make human lives easier can only do so by analysing the data that is presented to them. Another concurrent point is that with two of the described examples, the developers do not necessarily have a valid explanation for the results their algorithms have produced. The explosion of AI into every facet of our world has spurred critical discussion about the possible consequence of creating agents that are more intelligent than humans, agents that might have moral or legal status and what this means for the world. But an often-ignored issue is the oversight of the not so intelligent AI. Multitudes of tasks are now being delegated to smart machines that can compute things faster and process a lot more data than a human could. The problem is, these algorithms are influencing the world on large scale and sometimes we cannot explain how they come to their results and why they have come to their results. This generates a host of issues because although algorithms cannot be attributed with bias their unbiased nature does not negate their ill effects on society. The next section of this report will briefly touch on some of the key technological terms mentioned such as machine learning. The following section will further delve into the ethics and obligations of developers and the final section will conclude and summarise my thoughts on the matter.

# **Machine Learning**

Machine learning can be described as “the field of study that gives computers the ability to learn without being explicitly programmed” [4]. Machine learning algorithms are capable of analysing and learning from data then making predictions based on said data. They have become ubiquitous in today’s world and appear in search engines, stock trading and social media applications. Machine learning algorithms can be classified broadly into three categories of learning. The first is supervised learning; this involves providing the algorithm with input data and a desired output. The algorithm is then tasked with finding an appropriate function or pattern that can link the input to the output. Unsupervised learning involves the algorithm discovering and clustering data based on some identified correlating features. Reinforcement learning involves the algorithm learning through positive feedback; it learns by trying to maximise some reward it receives when positive actions are made. The approach taken by developers when trying to implement a machine learning algorithm varies according to the task necessary. Methods can take inspiration from biology like artificial neural networks (ANN) and evolutionary algorithms or statistics based methods such as decision trees or Bayesian networks could be used. Due to the complex nature of some of the methods used such as ANNs, it is hard to explain exactly how or why the algorithms are performing the way they are [5].

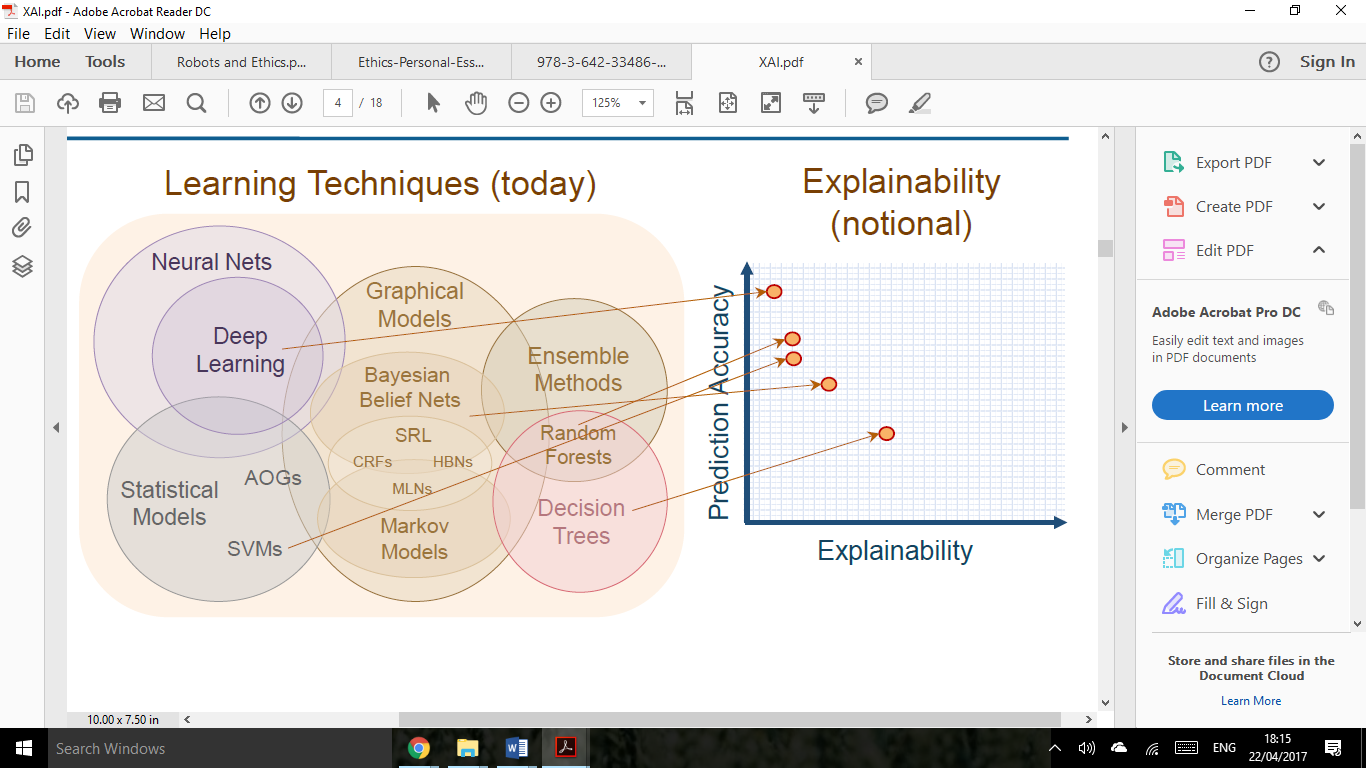


Figure : Performance VS Explainablity [6]

There seems to be an inverse relationship between the accuracy of the machine learning system and how easy its developers can explain what it is doing. Although this hasn’t prevented their use, deep learning which is a more complex and high performing application of ANNs has become a tech highlight ever since Google’s DeepMind. They made use of a reinforcement learning architecture which allowed them to defeat a world class player - Lee Sedol - in the board game Go. We have arrived at a point where we can employ algorithms that are extremely efficient at optimising goals and prediction tasks but these algorithms’ inner workings cannot be explained. Imagine a situation where patients are diagnosed extremely accurately but we can’t justify or provide reasoning for the diagnosis. What happens when the diagnosis is incorrect? How can we know if the system is behaving incorrectly?

# **Learning from what?**

Machine learning algorithms as demonstrated in the example of Tay can only base their inferences and models about the world from its current state. Given that the world has within it structural flaws that lead to inequality and discrimination, the algorithms will eventually learn to perpetuate and sometimes amplify these inequalities. Are developers then obligated to make sure these algorithms are properly vetted before being deployed into the world? James H Moor contends that a machine can be placed somewhere within these 4 categories as an ethical agent [7]:

* Ethical impact agents: these are agents whose actions have an ethical impact regardless of whether this impact is intended or not
* Implicit ethical agents: these are agents that have ethical considerations built into them, these are usually safety measures, for example a robot arm avoiding hitting people in its vicinity
* Explicit ethical agents: are agents that can reason about ethics and apply them to various situations
* Full ethical agent: agents that can make moral judgements and justify them. A human is an example of a full ethical agent.

In relation to these categories, it is uncontroversial to argue that robots or algorithms we deploy into the world should at least be implicit ethical agents. Their outcomes should be constrained by the developer to avoid unethical scenarios as best as they can. Achieving explicit ethical status is much harder. It might involve encoding ethical maxims into the machine’s programming which it can draw from and apply to appropriate situations with proper justification. As can be seen from the examples detailed in the introduction a lot of these machine learning algorithms do not meet the criteria of even an implicit ethical agent. With respect to the social cost of their actions, there does not seem to be any ethical consideration built into these algorithms before they are deployed into social situations. The data that these machines are learning from is allowing them to recreate a model of the world that we have now with all its flaws. The ubiquity of machine learning algorithms means they don’t just recreate the flaws of the world they have learnt from; they also magnify them and then learn from this further distorted world model. This creates a cycle that is self-perpetuating and can be hard to identify especially if the algorithms are operating in a complex indecipherable manner. The assumed knowledge about the results provided by amoral algorithms is that they are fairer and more trustworthy than any human could be.

This is further exasperated by the nature in which the data for these algorithms are collected. Data is usually collected with the aim of reducing the use of private and sensitive data like disability, race, gender, religion etc. This supposedly “blind” approach should prevent algorithms from using these protected characteristics to make decisions. Yet growing research is showing that machine learning algorithms can still manage to make indirect associations between some learned features and a protected characteristic. An example of this is the United States where the history of segregation means a person’s address can easily become a proxy for their race. This means we could have an algorithm that is unknowingly grouping people by their race, for example in a hiring process, without the explicit knowledge of the people making use of the algorithm’s decisions. Its discriminatory nature goes unnoticed. Some researchers have suggested the inclusion of protected characteristics into the data collection process in order to train the algorithm on identifying when it has created a proxy for these characteristics so it can better handle its own biases [10].

These sorts of issues have already begun to arise in the field of predictive policing where algorithms are being used to determine high risk areas that need police presence or the probability of recidivism for law breakers [8]. The major concerns for this are the possibilities of creating a self-fulfilling prophecy where areas that are monitored more lead to greater arrest rates which in turn leads to more police presence in these areas and ignorance of crimes in other under-policed areas. The ethical obligations of the developers need to be the proper consideration of their algorithms’ use; there is no big issue with the algorithm that YouTube employs to suggest videos to watch next but if the same algorithm is used for advertise jobs to users, it must avoid the pitfalls of indirect bias. The algorithms used by the police or companies for hiring purposes must be transparent and accountable. The rationale behind an algorithm’s decision must be well understood before it is used in these types of applications. However, having more transparent algorithms means they are more susceptible to manipulation. A balance between these two conflicting goals must be struck. The algorithms would need to become implicit ethical agents that try to limit the negative outcomes of their decisions. The Fairness, Accountability and Transparency in Machine Learning (FAT/ML) group is an established community of researchers concerned with tackling many of the issues raised in this report. Their work is essential to the further understanding of the unanticipated ethical issues that have risen in the wake of the machine learning revolution.

Some of the researchers [9] that have noted the possibilities of perpetuating already present inequalities with use of AI pushback against the idea of trying to remove bias within the algorithm itself by altering its perception of the world. They believe this would be generating an inaccurate world model which may cause further unforeseen complications in the algorithm’s operation. They argue, although the algorithm might be recreating an unequal world, the actions it takes can be contextualised to allow it to make more ethical decisions.

# **Conclusion**

Machine learning algorithms play a huge role in our everyday lives. They infer things about us and make important decisions about the information we are presented with and the information that is supressed from us. They possess an inherent quality of disinterestedness that allow them to be seen as unbiased and fair. Their opacity and complexity mean we are not privy to the processes which lead to their conclusions. They essentially become the gatekeepers who have the power to curate our lives without our direct knowledge or input. The ethical concerns they raise in terms of recreating and amplifying inequalities means we are at a critical point where we must assess how much power and control we are willing to entrust in their operation. Due to their general improvement of human life, efficiency and speed in a multitude of applications they are increasingly being deployed by corporate or state actors who can now use them in an unregulated manner without proper oversight and with the possible excuse of “the algorithm did it”. This is not to say that human decision makers are any less problematic, but society is already dedicating resources to combating implicit or explicit bias and discrimination. The same cannot be said yet for the systems that have been heralded as their future replacement.

# **References**

[1] Price, R. (2016) Microsoft is deleting its AI chatbot's incredibly racist tweets. *Business Insider UK* [online] 24 March. Available from: <http://uk.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=US&IR=T> [Accessed 20 April 2017].

[2] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92-112.

[3] Crawford, K. (2016) Artificial Intelligence’s White Guy Problem. *The New York Times* [online] 25 June. Available from: <https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0> [Accessed 20 April 2017].

[4] Munoz, A. (2014). Machine Learning and Optimization. *URL: https://www. cims. nyu. edu/~ munoz/files/ml\_optimization. pdf [accessed 2016-03-02]*

[5] Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge Handbook of Artificial Intelligence, 316-334.

[6] Gunning, D. Explainable Artificial Intelligence (XAI). URL*: http://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf pdf [accessed 2016-04-22]*

[7] Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21.

[8] Angwin, J., Larson, J., Surya, M., Kirchner, L. (2016) Machine Bias. *ProPublica* [online] 23 May. Available from: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [Accessed 20 April 2017].

[9] Caliskan-Islam, A., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *arXiv preprint arXiv:1608.07187*.

[10] Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183-201.