Deep Unlearning • Bias in AI Systems and Applications

Final Solution

Michelle L. ∘ *Arjun A. R.* ∘ *Azal A.* ∘ *Spandan S.* ∘ *Juhi S.* ∘ *Helen F.*

A multi-faceted approach to bias in AI involving a technical proposal, a solution for public interest, and an application of existing tools.

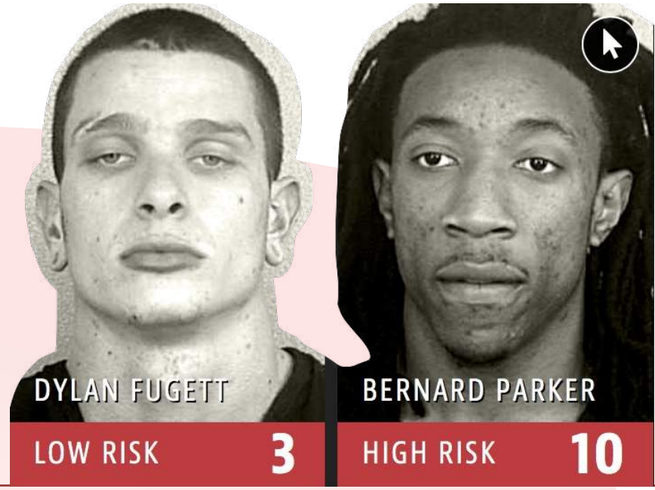
The Problem

Artificial intelligence (AI) is present in almost every aspect of daily life today. From automated job screening algorithms in business, to facial recognition technology in the criminal justice system, the impact of intelligent machines is extensive. Global use of AI grew by 270% from 2015 to 2019, and the market is predicted to reach $267 billion by 2027 (Lin, 2020). In fact, 85% of organizations across technological, financial, healthcare and government industries are currently evaluating or using AI in production (Magoulas and Swoyer, 2020).

However, as artificial intelligence spreads to even more industries like medicine and law, problems such as machine learning bias are likely to become more pervasive (Knight, 2017). As shown by recent controversies over bias in AI, e.g. the dispute over racial bias in the COMPAS prediction algorithm, artificial intelligence can exacerbate bias in unexpected ways.

Case study:

COMPAS recidivism prediction algorithm



Source: ProPublica

The Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, assesses the likelihood of arrested defendants to commit more crimes. It is widely used by judges and parole officers in US courts (Yong, 2018).

In 2016, an independent newsroom analysed the risk scores assigned to over 7,000 people arrested in Florida from 2013 to 2014, and checked to see how many were charged with new crimes over the next two years. They found significant racial disparities in the way that COMPAS predicted recidivism for black and white defendants, with notable bias against African Americans (ProPublica, 2016).

Three Core Issues

After reviewing current research in the challenge area, our team identified three categories of issues related to bias in AI.

1. Defining Fairness

In computer science, there are three formal fairness criteria: independence, separation and sufficiency. The Impossibility Theorem of Fairness shows that it is impossible for an algorithm to satisfy all potential fairness criteria at once (Zhong, 2020). **Instead, it is up to computer scientists to decide on what constitutes fairness for a specific machine learning system.** This is based on “user experience, cultural, social, historical, political, legal, and ethical considerations, several of which may have tradeoffs” (Google AI, n.d).

However, uncertainty and confusion over defining fairness for artificial intelligence is exacerbated by the fact that AI developers are not trained to be good ethicists (Ebert, 2020). According to PhD researcher Matthew Stewart, “Computer scientists, unlike doctors, are not necessarily trained to consider the ethical implications of their actions… Computer scientists are so far removed from data subjects that the implications on any one individual may be perceived as negligible and thus disregarded” (Stewart, 2020).

2. Human biases.

Automatic cognitive biases are essential to speed up decision-making in humans. However, these decisions come with "racial or social class categories or other unfair stereotypes" (*Susan Fiske and Shelley Taylor*, 2020). These human biases are difficult to eliminate. "Debiasing humans is harder than debiasing AI systems," claims Olga Russakovsky, an assistant professor in Princeton’s Department of Computer Science (Ghosh, 2021). **Society is full of these cognitive biases, which impact traditional human decision-making and leak into AI during training or user interaction, causing machines to reflect human prejudices (Manyika, Silberg and Presten, 2019).**

3. Biases in machine learning data and models.

In terms of bias in data, colorblindness and underrepresentation are core issues in AI which create discriminatory impacts on certain populations. For instance, minorities are less likely to have the time to establish a sound internet presence (due to making up a disproportionate amount of the lower class). This lack of representable online data can affect a job candidates’ access to jobs, as a lack of internet presence can act as a flag to an AI or HR department.

AI models are also blind to class intersectionality - the complex relationships between seemingly unrelated factors such as race, class and gender. A woman may be less likely to have higher paying jobs than a man, or a person of collor may be more likely to come from a lower tax bracket. Applying AI that does not understand nuances in data can produce unexpected discriminatory results.

Hypothesis

We predict that a comprehensive solution for bias in AI involves 1: a human-centered ethics awareness campaign, 2: a framework for applying AI to identify bias in machine and human decision-making, and 3: a technical model for minimising implicit bias in data. These components will effectively mitigate issues with defining fairness, inherent human biases and bias in technical models respectively.

Why use a humanist *and* technical approach?

It's worth putting in extra effort to improve processes behind AI's development and testing rather than just involving greater human supervision as a solution to bias because humans have inherent cognitive biases and data processing limitations.

Reducing cognitive strain to quicken decisions through automatic cognitive biases is a fundamental part of human function today. However, as Eric Colson states, "...'quick and almost unconscious' [doesn't] always mean optimal or even accurate" (Colson, 2019); these decisions come with the "racial or social class categories or other unfair stereotypes" (*Susan Fiske and Shelley Taylor*, 2020), which are dangerous when paired with AI's presence in important social and life events, like hiring (Zhang et al., 2019).

The autonomy of cognitive biases not only makes pre-training data-processing and developmental interactions with AI (*What it is and why*, 2018) more susceptible to translating peoples' biases into algorithms (Manyika et al., 2019), but also means removing them through ethics training alone can be resource-intensive and impractical.

On top of this, AI's ability to process high volumes of data makes algorithms less reliant on shortcuts that lead to prejudices in humans. Involving AI assistants in areas like job posting or application evaluation can even prevent or call out biased human decisions to adjust fairness for minorities and otherwise socially-disadvantaged groups in the employment process (Zhang et al., 2019). Increasing the independence of AI systems themselves reduces the effects of biases introduced through human interaction with algorithms, may be more practical than a purely humanist approach, and can check humans' cognitive shortcuts in ways not otherwise possible.

Solution Overview

1 Encourage awareness about AI ethics among new computer scientists.

We will create a website that democratizes skills in AI ethics by teaching and offering certifications about ethical awareness about bias in AI. In it, we will promote awareness and tools for assessing fairness. The curriculum will be based on guidelines from organisations such as the Association for Computing Machinery and existing certification courses from Google and the University of Helsinki.

2 Use AI bias to identify errors in humans' and machines' decision-making processes

As part of solution 1, we will encourage the use of existing tools by Google and IBM to recognise bias in the training process. We have explored the possibility of applying AI to existing data to detect implicit human biases in datasets and decisions in human systems, leading us to results like Adversarial debiasing. Additionally, we plan to do more research into the possibility of existing algorithms that can detect a biased algorithm, after it's development If feasible, we will outline a framework for applying AI in this way to real life situations.

3 Build a program to clean bias from data used for machine learning.

This program will adjust variables in a model and observe resultant outputs to detect variables which reflect implicit bias. It aims to fulfill counterfactual fairness, in which a model is considered fair if its prediction in the real world is the same as that in a counterfactual world where individuals belong to a different demographic group. This can be done through adversarial debiasing, in which two adversarial algorithms compete to predict and conceal sensitive data respectively. Adversarial debiasing can produce high-performing algorithms which are far less biased (Houser, 2019).

Web Course

Overview

Accessible Ethics Education

By relying on elements such as more manageably-sized lessons, an appealing visual style, interactive diagrams, etc., we can encourage more of the public to see ethical training as possible to achieve alongside current life commitments and as appealing as education in AI itself. This approach would help make a background in AI bias and ethics education more accessible and more integrated into a modern skillset--much like how companies like Codecademy, a programming education platform with a following of 45 million as of early 2020 (*Codecademy reaches 100,000*, 2020) and that relies on similar strategies, engage users to support the democratization of computer science education.

We believe that allowing this course to fit into existing classroom curricula would help place as much emphasis on AI ethics education as there is focus on teaching the technical aspects of artificial intelligence systems. By supplementing formal educationwith an approach focused on intuition, it may be possible to support longer-term retention and further maintain students' interest in an often-overlooked area of artificial intelligence knowledge.

This course's curriculum would involve areas such as the fundamental features that lead to bias in machine learning systems (e.g., the tendency of neural networks to weigh and link sensitive data in hidden manners); methods and measures of AI fairness as well as their shortcomings, such as fairness through unawareness and counterfactual fairness; tools like IBM's *Fairness 360* that exist to meet standards in AI fairness; the influence of AI systems and the impacts their objectivity can have; and the benefits less flawed AI decisions with regards to making up for humans' unconscious biases.

How does it address the problem?

Computer programmers are responsible for defining fairness, since there cannot be a universal definition of fairness for all AI applications - yet programmers are not well trained in the ethics side of technology. By educating programmers and the public about ethics and bias in artificial intelligence, we can promote better decision making in the design and implementation of AI.

Web Course

Practicality and Features

Why would a more informal course be practical?

The success of platforms like Codecademy--and other democratized-learning applications such as Brilliant.org, with “over 9 million” users[1] or DataCamp, with over 350,000 users[2]--gives us reason to believe this approach for helping others develop technical skills will fit well into the current landscape of skill-building. When asked about the potential practicality of our proposal, in addition, a university computer science major explained his view that a website of this variety would “...help make sure that new AI researchers are taking biases into account and help improve awareness,” a step in improving the perception that ethics, especially among those beginning in the artificial intelligence field, is not emphasized enough.

[1]Self reported
[2]Self reported
[3]More interface concept pages may be found here.

Using AI to Mitigate Bias

How can AI improve traditional human decision-making processes?

Proponents of AI have argued that artificial intelligence can be used in situations which require more impartial judgements, such as hiring processes where managers should make fair employment decisions. As explained in an article published by Harvard Business Review, "Machine learning systems disregard variables that do not accurately predict outcomes (in the data available to them). This is in contrast to humans, who may lie about or not even realize the factors that led them to, say, hire or disregard a particular job candidate..." (Manyika, Silberg and Presten, 2019).

However, using artificial intelligence in this manner has its drawbacks - in 2018, Amazon discovered that its AI recruiting tool showed unexpected bias against women applying for technical positions, as it reflected male dominance across the technology (Dastin, 2018). Rather, our team was drawn to the idea of taking advantage of AI's ability to pick up biases and prejudice easily, in order to expose human biases that may have gone unnoticed. In other words, our solution argues that we should use AI to detect biases in human and machine learning decisions. To test the feasibility of this concept, we attempted to apply existing tools for checking bias on sample data and sample models, including Google's *What-If Tool* and IB's *Fairness 360* toolkit.

Google's *What-If Tool*

First, we used Google's *What-If Tool* (WIT) to visually probe the behavior of a sample machine learning model across different inputs and for different machine learning fairness metrics. The WIT is a visual interface for probing the behaviour of machine learning models (Wexler, 2018). In our experiment, we applied the WIT to a public dataset[1] containing criminal history, demographics and COMPAS risk scores for defendants in Broward County[2]. COMPAS is a controversial algorithm used by US courts to predict recidivism in criminal defendants.

The WIT showed clear racial bias in a simple machine learning algorithm trained on the COMPAS data. Comparisons between inference scores showed how features such as race impacted predicted recidivism scores (1 = low risk). Categorising data by racial features also revealed that the algorithm disproportionately predicted low risk scores (blue) more for Caucasians than African-Americans. Thus, it seems feasible that we might use AI to detect biases in human and machine decision making.

Furthermore, we found the WIT easy to use for beginners in machine learning, as the tool was

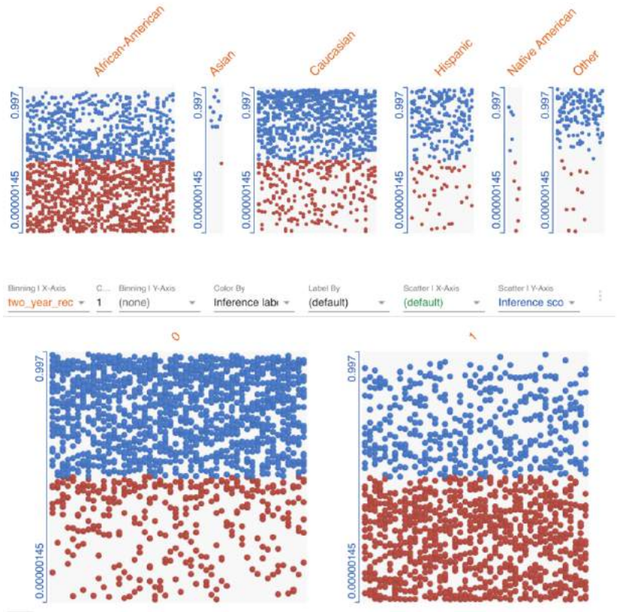
[1] *Study with a link to dataset used*

[2] *Please view our code here.*

Using AI to Mitigate Bias

Google's *What-If Tool*

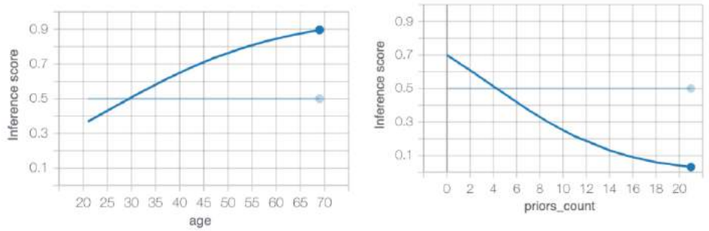
integrated with Google Colaboratory and had clear documentation. The flexibility of axis options allowed us to visually examine relationships between different factors and gain insight into where racial bias might appear. With further experimentation, we discovered that we could also manually edit the features of one datapoint to see how the recidivism prediction score shifted. We could also generate partial dependence plots to show the marginal effect of a feature on a model's predictions. The tool even suggested how threshold values should be altered to suit different definitions of fairness, e.g. demographic parity, equal opportunity, etc. Hence, the What-If Tool seems to be a feasible way for beginners to examine existing biases in algorithms.



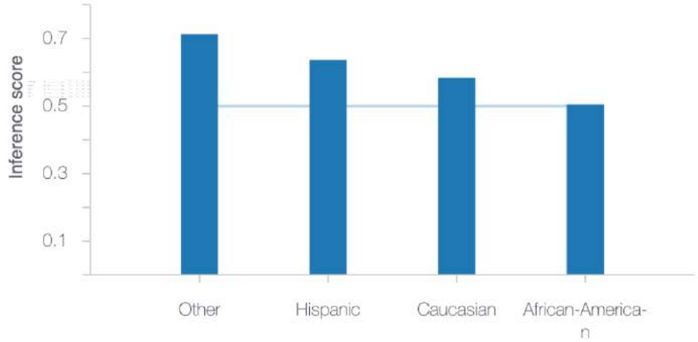
(Above) Shows clear racial bias in simple machine learning algorithms: disproportionately predicts low risk (blue) more for Caucasians than African-Americans

Run	Label	Score	Delta
2	0 (Medium or high risk)	0.593	↓ -0.065934
2	1 (Low risk)	0.407	↑ 0.065934
1	0 (Medium or high risk)	0.659	
1	1 (Low risk)	0.341	

(Left) You can manually edit the features of one datapoint (turning dials) to see how the recidivism prediction score shifts.



(Above) Partial Dependence Plots (PDP) show the marginal effect of a feature on a model's predictions.

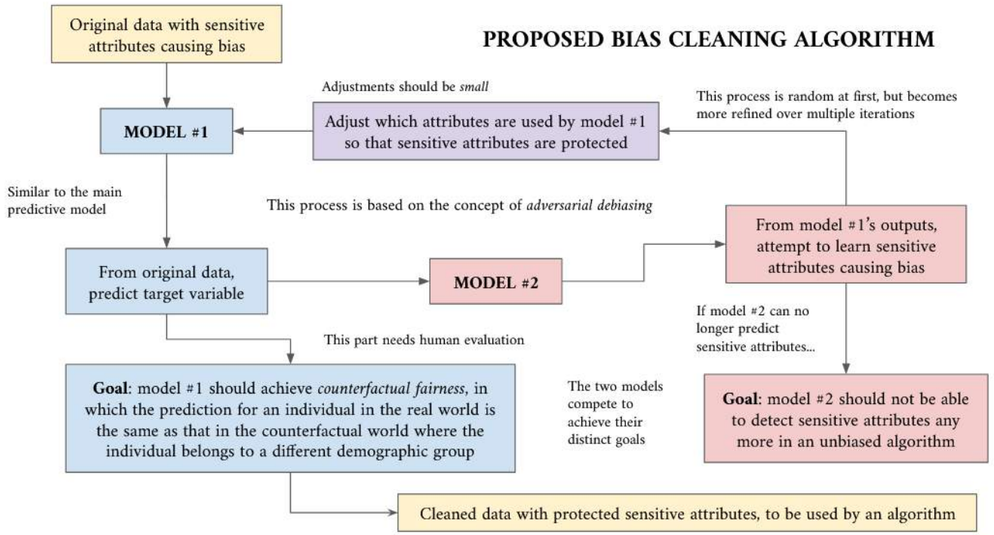


Proposed Bias-Cleaning Algorithm

Adversial Debiasing

One of the biggest issues with modern solutions to the issue of digitized bigotry is that of implementation. Algorithm usage already penetrates all aspects of modern society, from testing in sentencing and justice reform, to healthcare. All of these different situations would theoretically require personalized solutions for each, as they both have their bias manifest in different ways. It was this line of thinking then, that made our team realise we could never realistically come up with a large scale solution through modifying the existing algorithms. The issues are just so specialized and personal that it simply couldn't work. The only possible solution through this route was to create some sort of perfect algorithm that could read code and then evaluate issues with it. One of the largest issues with such an algorithm, is where it could deduce finality, or completeness. This problem is dubbed the Entscheidungsproblem, or the decision paradox. (Turing, 1937)

In reality, the bias introduced from algorithms is not an issue with the algorithms itself, but, as previously mentioned, it is an issue with the data. In a perfect unbiased world, algorithms would show no bias, because there is no bias in the data. There are no structures in society that indirectly affect various statistics about a person, and thus no indirect patterns an algorithm can observe. Our algorithmic solution is that of adverbial debiasing. A single algorithm, to remove the structural bias embedded in the data of a person. In the case of adversarial debiasing, we have two algorithms. One has a given dataset on a person (A) , the other tries to predict the protected identity of the person from the data (B). The goal of the first is to change the data slightly each time to remove power from the second, while the second trains to identify more in a loop, simulating the development of algorithmic biases in the long term. (Lemoine, 2018)



Proposed Bias-Cleaning Algorithm

Adversial Debiasing

Given that bias arises most from the underrepresentation and lack of consideration for certain nuances, making AI bias free is a challenge since their development is dependent on datasets. As such, the AI must be conscious of the fact that the data is likely skewed or biased in a certain way, which is the responsibility of the developer. In order to mitigate the effect of bias on data and thus AI, we suggest applying counterfactual fairness (Wu et.al , 2019) and generative modeling used in Generative Adversarial Nets (GANs) to generate a counterfactual world that is similar to the original data but ensuring that certain attributes are not the cause of a pattern in the data (Goodfellow et. al. 2014) in the proposed bias cleaning algorithm.

One of the largest issues found with adversarial debiasing, is that of speed. Randomized unweighting (turning of the dials) is simply not fast enough to make substantive change, and when it comes to more intersectional characteristics that appear in multiple dials simultaneously, it is difficult to teach the algorithm to hide, leading to a rapid point of diminishing returns as the algorithm changes dials that don't need independent change. Over time, the amount of debiasing done on a given dataset converges on some value, with a rapid point of diminishing returns. Additionally, adversarial debiasing has yet to have a functional implementation when it comes to biases stored in image recognition.

Overall however, the universibility of this solution simply cannot be ignored. Thus, a proposed implementation of this solution could have a user upload some sort of resume to a job hunting site (such as LinkedIn, Indeed, ZipRecruiter, etc), and on the website the data is then processed and debiased. This way, when companies receive data about the user, it has both already been processed, and decreases the needed work by companies to implement. This kind of implementation also allows for fast integration into the job application workflow, as only resume and job hunting sites would then need to implement the debiasing algorithm. (H.R, 2020)

Using AI to Mitigate Bias

IBM's *AI Fairness 360*

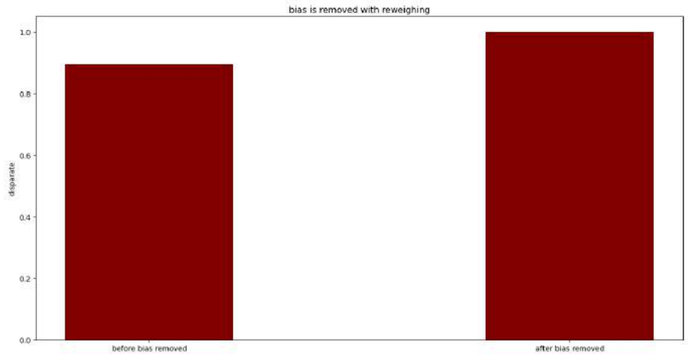
We also investigated IBM's *AI Fairness 360* tool to test out different bias mitigation algorithms and metrics on a sample dataset. *AI Fairness 360* is an open source toolkit for mitigating discrimination in machine learning models throughout the AI application lifecycle (Bellamy et al., 2019).

There are 3 ways of mitigating bias in AI algorithms given in the IBM *AI Fairness 360* module: pre-processing, in-processing and post-processing data. We have focused on using a preprocessing algorithm known as reweighing in this solution. Reweighing assigns a weight to each feature in the data to be applied before training a model with that data. This applied weight causes the model not to exhibit bias. After using a metric called disparate impact to measure the difference before and after using weighing on data, we came to a conclusion that reweighing did remove bias in the data.

We experimented with this tool to test for racial bias in the same COMPAS dataset. Applying Learning Fair Representations (LFR) and reweighing algorithms, we compared the effect of these preprocessing packages on recidivism predictions and disparate impact - the ratio of rate of favorable outcomes (low COMPAS scores below 12) for an unprivileged group (other races) to a privileged group (African Americans)[1].

Before applying the *AI Fairness 360* packages, the disparate impact was 0.89420 (5 d.p.). This was significantly improved to 1.00000 (5 d.p.) after applying LFR and reweighing. Granted, there are many alternative factors at play in this dataset, but our data clearly shows that AI Fairness 360 is a powerful tool for reducing potential sources of bias.

[1]Relevant code



This bar plot shows the disparate impact before and after reweighing the data provided. This causes the rate of favourable outcomes for the unprivileged group to equal the rate of favourable outcomes for the privileged group, causing the ratio to be 1.

Summary

How do you recognize and eliminate bias in raw data?

Along with promoting existing tools such as IBM's AI Fairness 360, a bias-cleaning algorithm based on adversarial debiasing and counterfactual fairness can be applied to clean data intended for machine learning models.

How do you define and measure fairness in the training model?

We cannot define a universal standard of fairness, but we can help and teach computer scientists to recognise the ethical consequences of fairness and bias in AI through a short online course.

How do you determine when a system is unbiased enough to be released?

We can use existing methods including Google's What-If Tool to test for bias in AI systems. Computer scientists (educated through our online course) can also use their best judgement to consider what level of bias is ethical for their specific model.

How can AI improve traditional human decision-making processes?

We can take advantage of AI's ability to absorb human prejudices and run bias-checking programs on models to detect subtle biases. Teaching programmers how to recognise bias in AI can also prevent the development of machine learning systems which execute unfair decisions (e.g., in the hiring process or in scoring recidivism).

How can ethical processes be established while mitigating bias in AI-based tools?

Educating computer scientists about bias and fairness in AI is the first step to bridging the gap between the humanities (ethics) and the sciences (technology). This will pave the way for more focus on ethical processes and fair systems in AI development.

Reflections and Future Development

While prototyping the web course, feedback from the target audience supported the viability of our solution. It would be easy and effective to develop a short online course on fairness and bias in AI, particularly as many computer programmers learn from online tutorials anyway. For future reference, we recognised that chunks of reading had to be interspersed with interactive diagrams and questions, in order to deliver information in bite-size pieces and maintain interest. To effectively promote an understanding of ethics in AI among computer science learners, our online course should emphasise interactive diagrams and peer-marked questions. Another idea for implementation would be to partner with existing online course providers, such as FreeCodeCamp, Codecademy or Coursera, to promote our educational content on bias.

Through experiments with existing AI to detect and mitigate bias in human decision making and machine learning, we verified the validity of our proposed solution. In particular, the What-If Tool seems to be a feasible way for beginners to examine existing biases in algorithms. Meanwhile, although less accessible to those without coding knowledge (i.e. general public), AI Fairness 360 is a comprehensive bias mitigation tool which should be promoted among computer scientists developing machine learning algorithms. Next, to create a more sustainable, ethical future regarding the implementation of AI, we should promote these easy-to-use bias checking systems to businesses and public organisations. This can be integrated into the first part of our solution, i.e. we can promote these tools to the public on our web course.

From prototyping our bias cleaning algorithm, it became clear that a technique inspired by adversarial debiasing would be easy to apply to classification and regression problems. However, there may be issues with speed, since randomised unweighting by model 2 may not be fast enough to make significant changes to the fairness of model 1. Complex relationships between intersectional characteristics would also hinder the efficacy of this model. Hence, the principles of our proposed model are feasible for simple machine learning models, but for more complex datasets, the algorithm would be slow and ineffective. Further research (possibly beyond the scope of this challenge) is needed to identify other bias-cleaning methods which could complement our proposed algorithm.

References

Acknowledgements: we would like to thank our mentor Abdul Rauf for his continuous support, and the Junior Academy for the opportunity to work on this project.