# CONTROLLING ALGORITHMIC BIASES IN AUTONOMOUS DRIVING CARS

Discrimination against minorities caused by biases integrated in self-driving car algorithms uncover multiple ethical issues in machine learning models

**Key Terms:**

Machine learning, Algorithmic biases, Input Data, Inequality, Algorithm, Artificial Intelligence, Autonomous Cars, Discrimination, Social Technology, Ethical Issues, Equity

Isabelle André

PHIL 250

University of British Columbia

Jasper Heaton, Jelena Markov

Date of Submission: 24 June 2019

***Introduction***

As the age of technology is in full swing, society has grown reliant on the use of physical and virtual machines. Where a human once operated, a machine is now accomplishing the same work in a more efficient and cost-effective manner. Due to the rise of the Internet and evolution of data storage, the proliferation of data "has now made machine learning a significant component of modern life" (Dickey et al., 2019, p.16), steering the world as we know it towards a modernized high-tech future revolving around computer algorithms and virtual reality. But are we truly moving towards the ideal technologically enhanced utopia that we seek? By analysing cause and effect of algorithmic bias, the grey zones associated with the equitability of machine learning models and training processes will be explored in order to determine on whom must the responsibility lie, thus allowing the generation of a fair solution.

***Machine Learning in Self-Driving Cars***

Machine learning is the main subset of artificial intelligence (AI) automating the development of analytical models. Controlled by algorithms, machine learning uses these models to analyse input data in order to learn patterns and forecast events using an iterative approach. The machine is then constantly exposed to large amounts of data in an automated process, allowing robust patterns to be recognized with actions made accordingly (Dickey et al., pp.16-17).

Despite still being in the testing stage of development, self-driving cars are the newest advancement in technology, rising in popularity among the wealthy. Machine learning algorithms have exponentially improved in the past decade to solve various challenges arising in autonomous car manufacturing. For instance, the car can receive information from its sensors, and process it through an application, allowing it to recognize the driver's speech and gestures, and even "[has] the capability to direct the car to a hospital if it notices that something is not right with the driver" (Savaram, 2017).

The machine learning used in AI algorithms can be classified as being unsupervised or supervised. Supervised algorithms are trained on pre-recorded dataset and constantly analyse patterns using regressions until they are able to minimize errors in their prediction, reaching an acceptable level of confidence in their forecast. Unsupervised algorithms develop their own relations from the dataset to detect the patterns and derive value from available data through clustering. Reinforcement algorithms are a third type of algorithms falling in between the latter types of learning, setting a certain target in each supervised trial with time-delayed and sparse labels (Dickey et al., p.17). Reinforcement learning has multiple practical applications addressing problems in AI, including the control engineering and operations of a self-driving car. A major task of machine learning algorithm is the "continuous rendering of surrounding environment and forecasting the changes that are possible to these surroundings" (Savaram, 2017). Some of these tasks include detecting objects on the road, its identification, recognition, and classification, and the object's localization and prediction of movement (Savaram, 2017).

***Case Study: Skin Colour Algorithm Detection***

Contrarily to many top car companies' beliefs, the implementation of fully autonomous cars into society may be a far more ambitious feat than previously assumed. According to a new study conducted by the Georgia Institute of Technology, the car's machine learning algorithms are more likely to misidentify certain demographic groups than others (Sigal, 2019). Individuals were first classified into groups based on skin tone using a method known as Fitzpatrick skin typing. The researchers then analysed the rate of success at which models identified the presence of people in a light-skinned group against those in the dark-skinned group. The error rate was on average five percent less accurate for the detection of the dark-skinned group. Controlling for variables such as time of day or the capture of obstructed pedestrian images exposed an even further disparity in the detection of various dark skin colour (Wilson et al., 2019, p.1). In conclusion, the standard models used for object detection trained on standard datasets was shown to exhibit lower accuracy on higher Fitzpatrick skin colour groups than lower Fitzpatrick skin colour groups, suggesting predictive inequity in the detection of individuals of darker skin tones.

***Disclosure of Training Model Dataset***

While the study shows valuable insight into the risks of algorithmic bias within automated cars, it was disclosed that the research tested "several models used by academic researches, trained on publicly available datasets" rather than "object-detection models actually being used by self-driving cars … [or] training datasets being used by autonomous vehicle manufacturers" (Sigal, 2019). As manufacturing companies refuse to publicly release their dataset for public scrutiny, researchers were compelled to use a restricted amount of relevant data, conceivably differing from the original data used to train the machine learning algorithms. Indeed, many onlookers have appealed for organizations "to be more transparent in terms of clearly spelling out the data collected, identifying which pieces of data are used in the algorithm, and disclosing how this data is weighted or used in the algorithm" (Kirkpatrick, 2016, p.17) to aid in pinpointing areas of discrimination that may not be obvious to those seemingly unaffected. The reluctancy to disclose available data exposes ethical issues regarding the transparency of machine learning and manufacturing companies, even more so as the problem becomes a matter of public interest.

***Algorithmic Bias and Human Bias***

The utopian vision of AI such as autonomous cars is that they would be operated objectively by systems unclouded by the prejudices and emotional factors of their human creators. However, a machine's desired objective outlook and its rational, unemotional decision-making process is precisely what becomes its worst flaw. Multiple studies like the one conducted by Georgia Institute of Technology have shown evidence that society instills their own prejudice and belief systems in the AI's automated decision-making systems (Sammy, 2019, p.45), also referred to as algorithmic bias. As algorithmic systems are trained with certain datasets, the AI will struggle recognizing unfamiliar instances when deployed if the data it is fed is

lacking in diversity. For instance, a model trained on taller adults may miss children and individuals of shorter stature. Other scenarios such as one where a person of darker skin colour is standing in front of a darker background, or pedestrians in wheelchairs crossing a street may not be well represented when training the AI detection model, resulting in a higher risk of error in the AI's detection system.

***Responsibility and Ethics***

There is still much ambiguity as to on whom does the responsibility of ensuring the equitability of the machine algorithms lies, exposing a predicament complicating the generation of a sustainable solution. Who within car manufacturing organizations should assume the responsibility of ensuring fair and unbiased decisions of machine learning algorithms? What can be done to remedy to the issue at hand to preserve the integrity of AI decision-making processes within self-driving cars?

As most clients lack basic understanding of the functioning of these algorithms, their inquires rarely touch upon the testing process itself, nor do they demand additional tests, reducing the odds of exposing any issues. As a result, dealers do not bother testing beyond their minimum set standards. The model's users should not be held accountable, as "they typically lack the expertise to evaluate [a machine learning] model" (Sammy, 2019, p.42). Furthermore, internal control principles dictate that "the person who creates a system cannot be impartial evaluators of that same system" (Sammy, 2019, p.42), suggesting that the model developer should not be held responsible for its AI's harmful decisions. After all, simply because a model is trained on biased data does not necessarily indicate that its programmer holds the same views, but rather that the sample data used in training was corrupted or incomplete. According to the Data Science Association's Professional Code of Conduct, developers have the duty to "protect the client from relying and making decisions based on bad or uncertain data quality" and "inform the client of all data science results and material facts known" (Data Science Association, 2019). Thus, once these biases become recognized, the party unwilling to rectify the problem would be held accountable for any further incidents.

***Controlling Algorithmic Biases***

Many observers take the problem and solution to be far simpler than it is, without realizing the true implications of some dilemmas computer scientists and engineers must face. How does one program an algorithm choosing whether a self-driving car holding a family should crash into another family or force steer its passengers onto oncoming traffic? It is easy to blame the creators of the machine, or the providers of the dataset which the machine algorithms are trained on, but as a senior research scientist at Google explains, "machine learning engineers care deeply about the measuring accuracy of their models" (Kirkpatrick, 2016, p.16). If the solution were so simple, perhaps there wouldn't be a problem to begin with. The common concept one would think of is to reconstruct the dataset to be more inclusive of minorities, but the problem is far more complex than simply increasing the sample data. To demand enough representation for every group is difficult when diversity can signify many things such as race, size, gender,

or nationality, resulting in a never-ending process. Furthermore, an entire population's data can be extremely hard to obtain or analyse, which is why it is often done with sample data, accurately representing a smaller portion of the population. No matter the amount of data, there will always be new exceptions surfacing, not fitting in with the rest of the clustered data, statistically referred to as outliers. We would therefore not only need to test for bias, but "constantly monitor for new types of bias as situations arise" (Wagner, 2019), as one can only measure bias in machine learning algorithms when the bias in question is known. Hence, it should be required that manufacturers conduct a series of explicit tests for known biases while also carefully monitoring for unexpected biases as they occur.

***Conclusion***

With scientists only beginning to comprehend the true dynamic capabilities of machine learning, society's laws and code of ethics fail to match the surge in resources and technological advances, uncovering new ethical issues and discrimination of minorities. While a heavier weighing in the dataset could help correct the bias to be more inclusive of minorities and outliers in the sample data, algorithms should constantly be monitored and tested, to ensure the fairness and consistency of decisions. Moreover, companies should have the duty to disclose their input data to maintain transparency in their decisions and algorithms, especially in such a situation where public safety may be involved. Doing so could allow observers to share insight and identify possible areas of discrimination that may not be apparent to manufacturers and developers. Such will also prove useful in educating the general public about the testing, decision making procedures, and risks of machine learning algorithms. As everyday tasks rely on big data and machine learning algorithms, one needs to learn to adapt in order to succeed in the new millennium. When operating with models that could possibly impact human safety, it is important to recognize the shortcomings of our current technical understanding, as "regulation at this point in time could easily do more harm than good" (Kirkpatrick, 2016, p.17).

References

Data Science Association. (2019) Data science code of professional conduct. *Data Science Association Code of Conduct*. Retrieved from http://www.datascienceassn.org/code-of-conduct.html

Dickey, G., Blake, S., Seaton, L. (2019) Machine learning in auditing. *CPA Journal (89)*(6), 16-21. Retrieved from http://web.a.ebscohost.com.ezproxy.library.ubc.ca/ehost/pdfviewer/pdfviewer?vid=22&sid=c938c9cd-ea04-42f4-ae7d-ef4d61d23753%40sessionmgr4006

Eubanks, Virginia. (2017). Automating Inequality. New York, NY: *St. Martin's Press.*

Johnson, Gabbrielle, M. (2017). Algorithmic bias: on the implicit biases of social technology. *UCLA Philosophy* Retrieved from https://philmachinelearning.files.wordpress.com/2018/02/gabbriellejohnson_algorithmic-bias.pdf

Kirkpatrick, Keith. (2016). Battling algorithmic bias. *Communications of the ACM, 59*(10), 16-17. Retrieved from http://web.a.ebscohost.com.ezproxy.library.ubc.ca/ehost/pdfviewer/pdfviewer?vid=14&sid=c938c9cd-ea04-42f4-ae7d-ef4d61d23753%40sessionmgr4006

Kropp, Andrea. (2018) Who's to blame for a biased algorithm? *Talent Daily*. Retrieved from https://www.cebglobal.com/talentdaily/whos-to-blame-for-a-biased-algorithm/

Martin, Kirsten. (2019). Designing ethical algorithms. *MIS Quarterly Executive, 18*(2), 129-142. Retrieved from http://web.a.ebscohost.com.ezproxy.library.ubc.ca/ehost/pdfviewer/pdfviewer?vid=13&sid=c938c9cd-ea04-42f4-ae7d-ef4d61d23753%40sessionmgr4006

Ravindra, Savaram. (2017). The machine learning algorithms used in self-driving cars. *KDnuggets, 17*(24). Retrieved from https://www.kdnuggets.com/2017/06/machine-learning-algorithms-used-self-driving-cars.html

Sammy, Allan. (2019). Bias in the machine. *Internal Auditor, 76*(3), 42-46. Retrieved from http://web.a.ebscohost.com.ezproxy.library.ubc.ca/ehost/pdfviewer/pdfviewer?vid=9&sid=c938c9cd-ea04-42f4-ae7d-ef4d61d23753%40sessionmgr4006

Samuel, Sigal. (2019). A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians. *Vox*. Retrieved from https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin

Wagner, Michael. (2019). Risk of AI bias in self-driving. *EET Asia*. Retrieved from https://www.eetasia.com/news/article/Risk-of-AI-Bias-in-Self-Driving

Wilson, B., Hoffman, J., Morgenstern, J., (2019). Predictive inequity in object detection. *Georgia Institute of Technology*. Retrieved from https://arxiv.org/pdf/1902.11097.pdf