# Artificial Intelligence and Ethics

Marcel Aguilar Garcia, Id: 20235620

January 8, 2022

## Part 1: Resource Creation

*Algorithmic bias and algorithmic fairness: The resource I have created is aimed for people with no knowledge of Machine Learning that want to understand how to know if they can trust machine learning models. I thought about it as a blog article for people with no technical knowledge. My goal is to communicate that, if the data or model are not public, and there are no regulations in place, it's hardly impossible to know if a system should be trusted.*

### Should I trust Machine Learning?

Machine Learning is impacting everyone's life. When you apply for a job or a personal loan in a bank, machine learning may be involved in the process. New products you have bought and your new hobbies may be influenced by ads that you have seen because of machine learning. Seeing how much it can impact us, we should ask ourselves, **should I trust Machine Learning?**

Let's start answering the question of **what is Machine Learning?** The dictionary defines Machine Learning as *a type of artificial intelligence in which computers use huge amounts of data to learn how to do tasks rather than being programmed to do them.* In other words, with machine learning, rather than computers doing tasks as they are explicitly programmed to do, they are learning how to do them through data. However, it is still up to the programmer to define when a task is performed well.

From this definition we can see that having a basic understanding of the data that has been used is essential. For example, we do not want to see a program using data from 1970 to predict if a candidate is going to be successful for a job position as this would reflect human biases and prejudices that would lead to machine learning mistakes and misinterpretations (such as gender inequality). Unfortunately, there are no regulations in place that require this sort of transparency and, many times, we need to trust it has been taken into consideration.

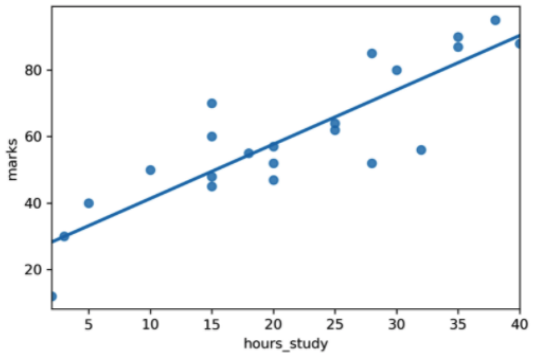
On top of that, we need to think that the algorithm will provide the probability of a person to be a 'successful candidate' for an interview but that this term is defined by the programmer. A 'successful candidate' could be considered the person that was hired after the interview process. However, it may be important to know the job performance for that person in the next years. We would not want to consider a person that was hired and did not perform well a 'successful candidate'.

While it is an important part, understanding the data that has been used does not provide the full picture. At this stage we could have a black box that provides random answers with the data that has been given. There are different methods that can be used to evaluate if the output of that box is correct. Let's imagine a program that reads pictures and has to predict if they contain dogs. If we would like to test how well this program performs, we could easily take fifty pictures of dogs and fifty pictures with no dog and count the guesses that the black box does, this number could be interpreted a percentage of how accurate it is. This black box is what we call 'model'.

At this stage, if I can see the data that the model has used and I am told that the accuracy of that model is high, I would definitely have more trust on the algorithm but still it is hard to believe if we are not able to see what is inside that black box.

This black box or model is an algorithm that uses statistical techniques to provide an answer. Models are not used just for prediction but as well to determine which data is relevant on a particular task. For example, a model that provides an estimation for the value of a house could help understand which fields are more important. How much more important is, e.g, the number of rooms on the house to the total size of the house? For this reason, it is important that we can understand the basics of how the algorithm provides the final answer.

Models that are simple are easier to understand, however they will usually not perform that well. As a result of this, when deciding which model to use for a particular problem, there is a balance that needs to be considered between simplicity and reliability. Simple models provide interpretability and transparency and, consequently, they are an important factor when trusting a system. One of the most used simple model is Linear Regression. Let's imagine we have the marks and hours of study of twenty students and, given the hours of study of new students, we would like to give an estimation of their marks. We can start plotting the hours of study and marks as seen in the plot below. The linear regression model associated to this data will be the line that passes closer through all these dots. If we are ask to provide an estimation for the mark of a student that has studied, e.g., 7 hours, we could give the value provided by the line (in this case we can see that is slightly below 40).




While mathematics may help, we do not need them to have basic understanding of this model. By looking at the visualisation, we can see that as hours of study increases, marks increases. We would be able to provide an estimation by ourselves and we would understand why we think it is close enough. The same may not be applicable if we increase the complexity of the model used.

Finally, let's try to answer the question **should I trust Machine Learning?** In general, private companies will not show the underlying data behind the model or the model itself. On top of that, at the moment there are no regulations or audits to assess these programs. Under these conditions, if I apply for a job position and I am told that there is a model that may discard or not my CV, I can not assess if the system is fair. As there are no regulations or audits, the trust I can have on the model will have to rely on the company policy, and the skills and knowledge of the specialist building this system.

## Part 2: Background literature review of the ethical issue

Recently, many companies have started using algorithmic decision-making systems in order to provide decisions that can impact consumer's life the lack of human interaction and the complexity for understanding how they work, have raised questions regarding the fairness of these systems [1] [2] [3]. The output of these algorithms can have an impact on important aspects of consumer's life, such as finance and health [2] [3]. For this reason, companies should always provide ways to ensure transparency and interpretability of these models [4][5]. There are many practices that could be applied to achieve these goals.

To start with, there is, in general, a balance between complexity and reliability of a machine learning algorithm [5][6]. When implementing a new machine learning system that could have a potential impact on consumer's life, it should be taken into account that the model does not only have to perform well but, should be able to be explained in simple and easy way. For this reason, rather than using a model to justify an answer, we should consider taking the decision of the algorithm, together with an understanding of why the algorithm has provided this answer. In this way, we should always think about the output of a model as the decision and the reasoning behind it [7]. Explainable AI (XAI) is the field that studies how to make the results of these algorithms understood by models [8]. While the focus of these literature review has been on algorithmic fairness, there are actually motivations for XAI. One of the motivations which is more relevant to our topic is 'Explain to justify', which allows consumers to make sure that there is no algorithm bias, especially when the algorithm does an unexpected decision. Another motivation is 'Explain to control' which can help to understand the system flaws and possible errors, by doing so, we can prevent incorrect behaviour before it happens. A third one, 'Explain to improve', helps to understand the advantages and limitations of a system, making it easier to be improved. Finally, 'Explain to discover', can be used to, not just use the algorithm to provide a decision but as well to gain insights from it [8].

While there are already some laws in place related to data privacy, there are almost no regulations that ensures consumers being treated fairly by algorithms [9] [10]. Corporations can use algorithms as a black box to consumers without having to show how the data and algorithms that they are using [9] [10]. With no regulations in place and no requirements to reveal the details of what has been used, there are no reasons for consumers to trust machine learning systems. This means that, the same way that Europe has adopted some measures to control data privacy (General Data Protection Regulation), there should be a similar way to ensure algorithm fairness, e.g., by auditing algorithms for bias. GDPR states that the consumers have a right to explanation that justifies the decision that has been done by the algorithm however this is an ongoing debate as the definition are vague and controversial [10] and, at the moment, they are not being regulated.

In conclusion, we have seen that in order to increase consumer's trust on algorithms, corporations need to provide interpretability by enforcing simplicity of the algorithm, and transparency for consumer's to understand how the decision has been done. While this is not always possible, companies should start to prioritise simplicity and interpretability over high performance models when a model can have a potential impact in consumer's life. A different way would be to legally regulate these algorithms, if the public can not have a clear understanding of the outputs of an AI algorithm, there should be external audits that inspect and assess the fairness of the algorithm to ensure that there is no algorithm bias.

# References


[1] R. J. Chen, T. Y. Chen, J. Lipkova, J. J. Wang, D. F. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithm fairness in ai for medicine and healthcare," *arXiv preprint arXiv:2110.00603*, 2021.

[2] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[3] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[4] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[5] Q. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *arXiv preprint arXiv:1802.00614*, 2018.

[6] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–52.

[7] K. Kirkpatrick, "Battling algorithmic bias: how do we ensure algorithms treat us fairly?" *Communications of the ACM*, vol. 59, no. 10, pp. 16–17, 2016.

[8] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[9] B. Casey, A. Farhangi, and R. Vogl, "Rethinking explainable machines: The gdpr's' right to explanation'debate and the rise of algorithmic audits in enterprise," *Berkeley Tech. LJ*, vol. 34, p. 143, 2019.

[10] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.