**“Intelligent Machines: Forget Killer Robots - Bias is the Real Danger in AI.”**

**- headline from a recent newspaper article.**

**Write up to 500 words discussing whether or not you agree with this headline regarding the imminent wide-spread application of AI and machine learning. You might, for example, focus on specific applications of machine learning and how biases might occur. Higher marks will be given for succinct, well thought through and well structured arguments. In particular, you should make reference to the concepts and methodologies presented in the lectures.**

The worry of deadly machines and killer robots has permeated public culture for years. Yet even as we make huge strides in machine learning, most experts agree we’re still far from a robot uprising [1]. Here I explore the far bigger danger of bias and how it will have huge consequences with the current wide-spread application of AI.

Increasingly within the US, predictive algorithms are used to compute risk scores for legal defendants. These scores are then used in an advisory setting to inform fundamental decisions on defendant sentencing. One of the widest used programs of this sort is COMPAS, which takes various factors from a defendant and then classifies their risk of recidivism. An article from ProPublica [2] explored the results from COMPAS and found that “Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism”. This is a key example of machine learning bias, as the algorithm is *biased*against black individuals because of the data used to train it. This can happen in many forms, either through highly unbalanced datasets, or simply biased data collection.

This highly worrying example isn’t to say that AI shouldn’t be used for life changing decisions, as in fact the human sentencing process contains many inherent biases too. The danger however lies with the fact that the biases here are hidden within a black box model. To the layman, these computational risk assessments offer a seemingly objective model of reality. The truth is however that statistical models aim only to provide simple approximations of reality. I believe the secretive nature of algorithms used are one of the main contributing factors to the danger of bias. The creators of COMPAS do not share specific calculations as it claims they are proprietary. One way to tackle this problem is to ensure that when AI applications are being used in the public domain there’s complete transparency for the data being used to train the model, and ultimately for how the model itself works.

The issue of the gap in understanding between machine learning practitioners and the general public exacerbates the problem further. In fact, the entire scope of knowledge for a machine learning engineer is not yet well defined. As an example, take Andrew Ng’s highly successful Coursera courses on Deep Learning. The modules are extremely practical and offer a great introduction to the subject. But if machine learning engineers aren’t encouraged to understand the underlying assumptions of these models, how will we expect the general public to?

The issue of bias is extremely worrying not only because of how easily hidden it is between ML practitioners, but also to the general public at large. In order to tackle the issue there needs to be more transparency in exactly how models are applied and which datasets are used to train them. This will ensure that there is more scrutiny over ML applications so as not to incur bias, an integral part of the scientific process.

Still need to:

* Need to talk about how transparency means more testing
* Relate it to stringent testing techniques
* And cross validation as mentioned in lectures