SIT799 Human Aligned Artificial Intelligence

Distinction Task 3.1: Mini literature review on bias and discrimination in AI

In this task I will be writing a literature review on domain of bias and discrimination in AI, specifically the quantification of bias.

Introduction:

Human biases are well-documented, from implicit association tests that demonstrate biases we may not even be aware of, to field experiments that demonstrate how much these biases can affect outcomes. Over the past few years, society has started to wrestle with just how much these human biases can make their way into artificial intelligence systems with harmful results. At a time when many companies are looking to deploy AI systems across their operations, being acutely aware of those risks and working to reduce them is an urgent priority. Bias can creep into algorithms in several ways. AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender, race, or sexual orientation are removed. For example Amazon stopped using a hiring algorithm after finding it favoured applicants based on words like “executed” or “captured” that were more commonly found on men’s resumes, for example. So is the reason why many researchers since decades are trying to find a suitable way to quantify the bias and reduce it as much as possible by unbiasing.

Key concepts, contrast and future scope:

In article [1] (Haussler 1988) the author gives a description of the notion that bias can be quantified in a way that enables to prove meaningful convergence property for learning algorithms, in this case a ML/AI algorithm. The author makes use

of a combinatorial parameter defined as the Vapnik-Chervonenkis dimension. The key concept upon applying this parameter is that the lower the dimension of the class of concepts considered by the learning algorithm, the stronger the bias. Further experimentation upon applying this measure has shown strong correlation with the learning performance. The author further provides numerous hypothesis and the proof of concept of the same. The author further concludes the study by mentioning the fact that the analysis used is still inadequate on several accounts of learning algorithm. The few drawbacks of the study are that there was no consideration of possible misclassifications in the training samples and the study also doesn't provide a clear way of how algorithms can be modified to tolerate such misclassifications. Moreover, the methodology proposed by the author may not be appropriate one for incremental learning algorithms i.e. algorithms that maintain a working hypothesis and update this hypothesis as new examples are received.

In article [2] (Turney 1995) the author talks about the fact and concerns with the impact of bias on predictive accuracy. And mentions about the possibilities of other how other factors also play a role in evaluation of bias such as the factor of stability of the algorithm i.e. repeatability of results. The authors also distinguish the category of bias generally as of two types namely Exclusive/representational bias and preferential/procedural bias and explains further on these types. Representational bias is typically a form of exclusive bias, since constraining the representation language means that certain concepts cannot be considered, since they cannot be expressed. Procedural bias is typically a form of preferential bias. The author also mentions the measure used by author in article [1] (Haussler 1988) the VC (Vapnik-Chervonenkis) dimension is a measure of the strength of an exclusive bias (Vapnik, 1982; Haussler, 1988). The author introduces another measure name Schaffer's as a component in

his definition of a measure of the strength of preferential bias. The idea behind Schaffer's definition is that a learner with a strong preferential bias toward f1 over f2 will only learn f2 when f1 is substantially more accurate than f1. The author further proposes various theorems followed by the proof of the same. Few of the drawbacks which are to be future scope of research suggested by the author are about stability and its relation to accuracy and bias are as follow. How do distinct learning algorithms, such as neural networks and decision trees, compare with respect to stability? (Turney 1995) How can a learning algorithm adjust the strength and type of its bias to suit the data it faces, to optimize the correctness of its bias? When measuring the correctness of a bias, how can we combine the criteria of accuracy, stability, cost, and complexity? How do we handle the case where the set of classes, C, is infinite? (Turney 1995) That is, how do we measure stability when learning to fit a curve, rather than learning to classify? (Turney 1995).

Conclusion:

Any examination of bias in AI needs to recognize the fact that these biases mainly stem from humans' inherent biases. The models and systems we create, and train reflect ourselves. Few studies have assessed the effects of gender bias in speech with respect to emotion — and emotion AI is starting to play a more prominent role in the future of work, marketing, and almost every industry you can think of. So how do we solve this problem at hand, there are few practices currently applied for ML/AI algorithms to avoid gender bias like ensuring diversity in the training samples (e.g. use roughly as many female audio samples as males in your training data), encouraging machine-learning teams to measure accuracy levels separately for different demographic categories and to identify when one category is being treated unfavourably, Solve for unfairness by collecting more training data associated with sensitive groups. From there, apply modern machine learning de-biasing

techniques that offer ways to penalize not just for errors in recognizing the primary variable, but that also have additional penalties for producing unfairness.

Although examining these causes and solutions is an important first step, there are still many open questions to be answered. Beyond machine-learning training, the industry needs to develop more holistic approaches that address the three main causes of bias, as outlined above. Additionally, future research should consider data with a broader representation of gender variants, such as transgender, non-binary, etc., to help expand our understanding of how to handle expanding diversity.

We have an obligation to create technology that is effective and fair for everyone. I believe the benefits of AI will outweigh the risks if we can address them collectively. It’s up to all practitioners and leaders in the field to collaborate, research, and develop solutions that reduce bias in AI for all.

References:

1. **Haussler, D 1988, "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework", *Artificial Intelligence*, vol. 36, no. 2, pp. 177-221.**

2. **"Quantifying the Risk of AI Bias" 2020, retrieved 11 August 2020, <https://levelup.gitconnected.com/quantifying-the-risk-of-ai-bias-998a5542a5e0>.**
3. **Turney, P 1995, *Machine Learning*, vol. 20, no. 1/2, pp. 23-33.**
4. **Violago, V & Quevada, N 2018, 'AI: The Issue of Bias', Managing Intellectual Property, vol. 277, pp. 32–36, viewed 11 August 2020, https://search.ebscohost.com/login.aspx?direct=true&db=edshol&AN=edshol.hein.journals.manintpr277.14&authtype=sso&custid=deakin&site=eds-live&scope=site**
5. **Britt, P 2020, 'Overcoming Bias Requires an AI Reboot Biases in artificial intelligence can be mitigated with multiple programmers', Speech Technology Magazine, no. 2, p. 14, viewed 11 August 2020, <https://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.625863761&authtype=sso&custid=deakin&site=eds-live&scope=site>.**