Project 2: Ethics in AI

Class: MO434 - Deep Learning
Institute of Computing - UNICAMP

Group members:
Felipe Marinho Tavares **RA**: 265680
Matheus de Souza Ataide **RA**: 147375

1 Definition of Ethics in AI

Ethics in AI is the concern to ensure that agents and system containing elements of artificial intelligence behave morally or as though moral. With the notorious increase in the use of AI in recent years, many problems are arising related to ethics in these new intelligent systems. Some of the main topics of debate in this field are: Privacy & Surveillance, Manipulation of Behaviour, Bias in Decision Systems, Human-Robot Interaction and Employment.

Privacy issues are usually caused by using personal data in modern AI systems. Controlling who collects which data, and who has access, is a hard task in the digital world. Many new AI technologies amplify the known issues. For example, face recognition in photos and videos allows identification and thus profiling and searching for individuals. Furthermore, these collected data can be used by algorithms to target individuals or small groups with just the kind of input that is likely to influence these particular individuals, this process is known as manipulation of behaviour.

Another very common cause of ethics related problems is the bias applied by the creator of an AI model. By using training datasets containing a biased distribution of samples, the resulting model might perform discriminatory inferences on its predictions. Also, as the creator chooses when to stop training and how to tune the AI model, bias might be introduced in this process in a negative way, possibly causing discriminatory behaviors in the model. A wide range of different bias can be introduced in a model, some examples are: Historical Bias, Representation Bias, Measurement Bias, Evaluation Bias, Aggregation Bias, Population Bias, Sampling Bias, Content Production Bias, Temporal Bias, Popularity Bias, Observer Bias, and Funding Bias.

2 Recent news involving Ethics in AI

In May of 2020, Nick Statt published in *The Verge*, a technology news website, a report entitled "ACLU sues facial recognition firm Clearview AI, calling it a 'nightmare scenario' for privacy". It is about the case of Clearview AI, a company that uses artificial intelligence technology. This company was sued by the American Civil Liberties Union because of its facial recognition system for violation of the Illinois Biometric Information Privacy Act (BIPA), alleging the company illegally collected and stored data on Illinois citizens without their knowledge or consent and then sold access to its technology to law enforcement and private companies. This happened in May of 2020. The complete report can be found at https://www.theverge.com/2020/5/28/21273388/aclu-clearview-ai-lawsuit-facial-recognition-database-illinois-biometric-laws

3 Paper about a solution to an Ethics in AI problem

Wang et al. [1] publicated work in CVPR 2020, named Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation, proposes a benchmark to evaluate the biases learned by algorithms trained with human images, and a training method to reduce said biases. Table 1 shows the paper detailed information.

Name	Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation
Authors	Zeyu Wang, Klint Qiname, Ioannis Christos Karaksozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky
University	Princeton University
Publication date (online)	2 April 2020
Google Scholar citations	14

Table 1: Wang et al. [1]

By works such as from Zhao et al. [2] it's shown that trained algorithms learn to some degree demographics from it's human data such as gender.

References

1 WANG, Z. et al. Towards fairness in visual recognition: Effective strategies for bias mitigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 8919–8928. Google Scholar citations: 14. url: <https://arxiv.org/pdf/1911.11834.pdf>.

2 BOLUKBASI, T. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, v. 29, p. 4349–4357, 2016.