**Name:** Eamonn Herlihy **Student ID:** 14155605

**Module:** IN5103 - Risk, Ethics, Governance and AI

**Assignment:** Week 3: Investigating AI Society, Commercial AI Products, Citizen/Users and a Risk/Ethics/Governance Ecosystem

**Task - Construct two case studies regarding emerging AI innovation**

**Case Study One: Bank that Doesn’t ‘See’ Race or Gender**

Stephen has recently finished a master's in AI and was very fortunate to land a lucrative job at start-up New Bank. New Bank is a FinTech company that gains a competitive advantage from its extensive use of AI solutions throughout its business. Stephen’s first assignment is to work on New Bank’s Credit Card product suite, in particular the AI techniques used to derive the line of credit offered to customers. This AI calculator is a source of great pride for New Bank and the product team has received many internal awards for their continued success. While getting up to speed with the models in place, Stephen quickly notices that white males tend to receive larger lines of credit, even when compared to those who share many of the same attributes (for example, job description, salary range, etc). In the brief, Stephen remembers that New Bank explicitly stated that the model was programmed to be ‘blind’ to gender and race, meaning that gender and race were not used as input features to the model. Stephen still felt uneasy about the outcomes he was observing. He remembered from an ethics course he had taken that there is an abundance of research showing how correlated inputs (or proxies) can lead to unwanted biases in complex algorithms [\*see source]. Stephen is still on a probationary period as he recently joined New Bank, has substantial student debt, and does not want to ruffle any feathers by criticising this innovation that is New Banks' crowning jewel.

Put yourself in Stephen’s position, should he raise concerns about this potential bias? Is there a moral obligation for New Bank to critically review this issue and rectify any inequalities, even at the expense of reduced accuracy in their coveted credit line calculator?

*Source:*

O'neil, C., 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

<https://www.youtube.com/watch?v=gdCJYsKlX_Y>

**Case Study Two: Facial Recognition in Car Parks**

Stephanie has recently finished a Ph.D. in computer vision and is currently trying to decide on her next career move. She has two offers from big tech companies offering sizeable income packages but always dreamt of running her own business. Several months back, while in a large shopping mall in Dubai, Stephanie had forgotten where she parked her car. What resulted was several hours of painstaking searching before the eventual success. It was just then, the idea dawned on her – ‘why can’t we use computer vision and facial recognition to track where I parked my car?’. Stephanie envisioned that using computer vision to grab the number plate and facial recognition to identify the driver, a system could be set in place that allowed a customer to use their face to find exactly where they had parked their car. This way Stephanie would not have lost hours during her Dubai trip to this tedious task. Stephanie has already talked with potential investors who seem very interested in the data collection and less so with the actual intended use case. She is somewhat troubled by this. Stephanie is afraid that this on-the-surface novel and harmless use case could cause more harm than good, if the will of hyper-commercial profit driven actors take hold.

Analyse this scenario in terms of the potential benefits and risk. In the implementation of such a solution, where will Stephanie need to be very careful from an ethical AI point of view? Can you identify other seemingly harmless products that have turned out to be questionable from an ethical point of view?

*Source:*

<https://fortune.com/2021/08/24/eye-on-a-i-data-privacy-unethical-kpmg-survey/>

**Respond: Analyse one case study from your peers**

**Case Study to Review:**

*(Brian Wood’s via Forums)*

Karen has recently graduated from a post-graduate programme in Artificial Intelligence and has been offered a position with a tech firm working in the health sector. Karen has a background in health sciences and thus is a perfect fit for this group as they transition into the area of health screening. The company has recently been offered a lucrative contract with a national health provider to deploy an AI system for ocular screening of patients with diabetes.

Diabetic patients have photos taken of the back of their eye on an annual basis. Traditionally this photo would be examined by a trained technician who then decided whether:

1. the patient could continue with their annual check if no changes associated with diabetes or only mild changes were noted;
2. the patient needed to be brough back for a repeat photo at a sooner interval if more pronounced changes were noted;
3. or whether they should be sent to an eye specialist for treatment if severe changes were noted.

The new AI system developed has been shown to be very accurate at picking up signs of diabetic retinopathy and triaging patients with follow up appointments of varying intervals depending on severity or else referrals in to see and eye specialist if severe disease is noted. The health service is very interested in bringing this system in as they anticipate significant cost-savings with the reduction in manpower this system allows.

Karen is tasked with working on this project. The system performs extremely well on all training sets. However, Karen notices that the training data have been taken from a predominantly Caucasian population in a different jurisdiction. She also notes that the majority of the images in the training set have been taken from highly labelled photos from randomised controlled trials that used the same high-quality camera across all trial centres. She knows that the cameras that will be used in the new screening centres are not uniform and that there is more variance in the ethnicities of the area in which this new equipment will be deployed.

What ethical concerns might Karen have?

**Response:**

This solution has the potential to provide obvious societal benefits. The use of AI in the medical fields has been growing rapidly in recent years and the literature suggests AI technology can indeed out-perform the human. For example, Yu and Lehman (2018) describe the machine learning approach that has the promise to significantly increase the accuracy of breast cancer classification (benign/malignant) and argues that this could drastically reduce the number of ill-required surgery. Although the benefits seem obvious i.e., increased diagnosis accuracy, better treatment strategies and reduced workloads for staff etc., the risks are more subtle.

Using the Beauchamp and Childress (2001) framework for “common morality”, there are several risks and ethical questions/concerns that need to be addressed. The case study suggests that Karen is quite concerned about the ‘Justice’ (distribute health resources fairly) and ‘Nonmaleficence’ (first do no harm) principles. The non-representative training data used to create the AI runs the risk that the outcomes will be unequitable i.e., as predominantly Caucasians were used to test the system, it is likely that this population will receive more favourable outcomes. Karen also notes that there will likely be a mismatch in camera quality between the development (training/testing) and the implementation. This will run the risk that those screening centres who use lower/different quality imaging may achieve poor results and thus putting patients at risk (‘Nonmaleficence’). Moreover, if low-income geographies happen to have lower quality camera’s, this could also have further repercussions to the ‘Justice’ principle.

Additional risk/concerns remain i.e.,

- so-called ‘black box’ models can violate the ‘respect for autonomy’ principle

- individuals’ data, data security, and consent need to be considered

- potential AI failure and accountability

- automation complacency

- etc.

Source:

Beauchamp, T.L. and Childress, J.F., 2001. Principles of biomedical ethics. Oxford University Press, USA.

Yu, L. and Lehman, C.D., 2018. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology, 286(3), pp.810-818.