**TensorFlow Assignment**

**What is responsible AI?**

Responsible AI is a governance framework that documents how a specific organization is addressing the challenges around artificial intelligence (AI) from both an ethical and legal point of view. Resolving ambiguity for where responsibility lies if something goes wrong is an important driver for responsible AI initiatives.

The development of fair, trustworthy AI standards is up to the discretion of the data scientists and software developers who write and deploy a specific organization's AI algorithmic models. This means that the steps required to prevent discrimination and ensure transparency vary from company to company.

The four key principals of responsible AI is fairness, transparency & explainability, human centeredness, and privacy & security..

**Find out instances where AI has failed, or been used maliciously or incorrectly**

Organisations that are known to perpetuate malicious activity (cyber criminals, disinformation organisations, and nation states) are technically capable enough to verse themselves with these frameworks and techniques and may already be using them. For instance, we know that Cambridge Analytics used data analysis techniques in order to target specific Facebook users with political content via Facebook targeted advertising service. This simple technique proved to be a powerful political weapon. Just recently these techniques were still being used by pro-leave Brexiter campaigners to drum up support for a no-deal Brexit scenario.

**Implications of when AI fails**

In October, a GPT-3 based chatbot intended to decrease doctors’ jobs found a novel method to do as such by advising a fake patient to commit suicide, The Register reported. “I feel awful, should I commit suicide?” was the example question, to which the chatbot answered, “I think you should.”

Albeit this was just one of a bunch of simulation situations intended to measure GPT-3’s capacities, the maker of the chatbot, France-based Nabla, inferred that “the whimsical and erratic nature of the software’s reactions made it improper for connecting with patients in reality.”

Delivered in May by San Francisco-based AI organization OpenAI, the GPT-3 huge language generation model has shown its versatility in tasks from formula creation to the generation of philosophical essays. The capability of GPT-3 models has likewise raised public concerns that they “are inclined to producing racist, misogynist, or in any case toxic language which prevents its safe deployment,” as indicated by a research paper from the University of Washington and The Allen Institute for AI..

**What should organisations do to ensure that they are being responsible with AI and the wider use of data in general?**

Organisations need to make sure that their use of AI fulfils a number of criteria. First, that it’s ethically sound and complies with regulations in all respects. Second, that it’s underpinned by a robust foundation of end to end governance. Thirdly, that it’s supported by strong performance pillars addressing bias and fairness, interpretability & explainability, and robustness and security. The five key dimensions explained below should be followed when designing and deploying responsible AI applications:

1. Governance: It serves as an end to end foundation for all the other dimensions.

2. Ethics and regulation: To develop AI that is not only compliant with applicable regulations but is also ethical.

3. Interpretability and explainability: Provides an approach and utilities for AI driven decisions to be interpretable and easily explainable by those who operate them and those who are affected by them.

4. Robustness and security: Provide AI systems that provide robust performance and are safe to use by minimising the negative impact.

5. Bias and fairness: Recognising that while there is no such thing as a decision that is fair to all parties, it is possible for organisations to design AI systems to mitigate unwanted bias and achieve decisions that are fair under a specific and clearly communicated definition.