Week 3 - SGTA Task

1. *Johnson (2021) develops and discusses the 'proxy problem' in their account of algorithmic bias. Please characterise the problem and give an example of its occurrence.*

The 'proxy problem', as described by Gabbrielle M. Johnson (2021) arises when innocuous or seemingly unrelated attributes in algorithmic decision making serve as 'proxies' for socially sensitive attributes such as race, gender, and religion. Even when these sensitive attributes are deliberately removed from the dataset, the machine learning algorithm can still infer these traits indirectly through correlated proxy features. This happens as machine learning models infer patterns in data that may not be immediately apparent to researchers and engineers, leading to algorithmic biases that reflect real world biases against marginalized groups.

Employment algorithms used for screening job applicants are a prominent example of this problem. Even with anonymized data, machine learning algorithms trained on historical hiring data that reflects a preference for male candidates may reflect these biases without being explicitly told to. This can happen even when gender data has been removed from the dataset, as the AI may use other proxy features such as schools attended, language used in resumes, or work history to form biases that prioritize male applicants.

2. *What exactly is the 'value neutrality thesis' as discussed by Fazelpour and Danks (2021)? Does it accurately capture the relationship between algorithmic decision- making and the reality of our socially and politically shaped lifeworld?*

The 'value neutrality thesis' is the idea that since a machine learning algorithm's output is determined by the training data use, that it cannot hold values or beliefs. Fazelpour and Danks provide an example in the form of an analogy, "one would typically not say that a hammer is biased, though it can be used in biased ways." (Fazelpour & Danks, 2021).

Fazelpour and Danks also discuss the arguments against the value neutrality thesis, and how it fails to accurately account for reality and the socio-political climate. Firstly, as a machine learning model enables certain decisions or actions, and as such implicitly infer that these decisions and capabilities are important. And secondly, that models are evaluated based on their adherence to a standard, aiming to maximize some metric. As such, the model is more likely to prioritize responses that maximize the metric it is being measured against and potentially neglect other metrics that its creators or users deem less important. Overall, the value neutrality thesis describes an ideal scenario, in which algorithms are an unbiased tool used for purely data-driven decision making. This fails to account for the methods of which the model is created, optimized and implemented.