**Notes:**

**Ethical dilemma of self-driving cars:**

* AI can only make decisions and cannot act instinctively
* May be programmed to prioritise the life of the consumer and not consider external factors
* May be seen as homicide as the algorithm is predetermined

**Algorithmic bias in AI and Understanding AI ethics:**

* Algorithmic bias may allow for AI to produce unfair or discriminatory outcomes
* May be caused by bad training data or misclassified data -> causes a feedback loop that reinforces the bias
* Biases that arise through programming or proxy data
* Application of results from AI may be biased
* Examples – recruitment: AI hiring program can be biased towards male applicants as majority of past resumes were from males, finance: AI looks at past data, so has demographic biases and marginalises minorities
* Prevention: representative data that is collected through stratified sampling so the results are more inclusive, bias detection through impact assessments and causation tests, transparent AI to fully understand why the outcomes are why they are, inclusive AI development team

**Case study:**

<https://www.timesnownews.com/technology-science/racist-ai-stuns-all-transforms-asian-mit-grad-into-a-white-woman-for-linkedin-article-102301138>

Shows AI bias due to biased training data, as when a woman asked for her LinkedIn profile picture to be made more professional, it changed her race into a white woman, showing racial bias in the training data, as it was unable to identify the woman’s race and keep it constant.

**Reflection:**

Overall, today I have learnt that AI is incredibly problematic due to biases that arise in a multitude of areas. I have come to realise that the main way to mitigate these biases from emerging is to use unbiased and up to date training data, as many AI models are trained using old-fashioned and outdated information, which leads to these situations in which certain groups are marginalised through AI.