Critically discuss key parts of AI and how they can contribute to generating biased outputs.

Artificial intelligence (AI) is ever present in the modern world and is the process of teaching computers how to recognise patterns, as well as understand language in order to solve human problems. This is done using algorithms which allow the AI to analyse large datasets and learn to be able to predict future outcomes. It is a prominent part of our everyday lives which is present in the form of virtual assistants, such as Apple's Siri and Amazon's Alexa, or programmes which are designed to predict shows you may want to watch, Netflix recommendations, and products you might want to purchase to name a couple examples. Francesca Rossi – a professor in computer science and the AI ethics global leader at IBM – categorises AI into two key areas of research, "one is based on rules, logic, and symbols; it is explainable; and it always finds a correct solution for a given problem, if that problem has been correctly specified" and the other is "based on examples, data analysis, and correlation" (Rossi, 2019). The two areas which Rossi is describing are symbolic AI and machine learning. Symbolic AI algorithms are given specific rules or shown specific symbols in order to understand and process data, whereas machine learning involves exposing AI to large datasets, compiled of images, text, or audio, with the goal of looking for correlations within the data. There are many parts which come together to form AI such as data, datasets,

algorithms, models and training however the key areas I will be focusing on in order to answer this question will be the data that goes into datasets, the large datasets which are used to teach AI, and the models which outline the patterns which are found within the datasets.

In order to discuss how components of AI may contribute to bias, we first need to understand what bias means. Dictionaries understand bias to be prejudice against a person or a group in a way which is deemed unfair. This is to mean that because of a preconceived idea, one person or group receive different treatment based on stereotypes. Bias within machine learning stems from human intervention when developing algorithms, creating datasets and training AI. For an AI system to be categorised as biased, Yeung et al. came up with the parameters that it must “(1) consistently produce disparate or disproportional outcomes for different groups of people and (2) the disparate impacts are not commensurate with what might be expected for people in the affected groups given their relative proportion of the population” (Yeung et al., 2021). This is to mean that an AI system is considered bias if it both consistently creates different results for different groups of people and if the information it provides is out of line with what is expected from such groups. Another term that is mentioned a lot within the discussion of AI is language models (LM). In their article on whether language models can be too big,

Bender et al define language models as “systems which are trained on string prediction tasks: that is predicting the likelihood of a token given either its preceding context or its surrounding context” (Bender et al, 2021). In regards to LMs, tokens are the information that LMs are fed and the information that they generate and they can be in the form of single letters, parts of words, or whole words. For example, the word ‘impossibility’ does not get processed by an LM as it reads, but instead gets split up into the tokens ‘im’ and ‘possibility’. Language models are taught using tokens rather than being fed individual words in order to improve efficiency and to allow them to recognise patterns. In this case, LMs will learn to understand that when the token ‘im’ comes before a word, the meaning of the word is reversed – something possible becomes impossible or something probable becomes improbable. Chatbots such as GPT-3, Claude, and BERT are all examples of what an LM is and are taught using these tokens in order to read large datasets; one area that can develop bias within these forms of AI.

So why is bias within AI an issue? Now that we have looked at what bias within AI looks like, we need to understand the issues that in can cause. The main issues I have selected that AI bias can cause are unfair treatment, inaccurate decisions, loss of trust, and ethical problems. Unfair treatment of already marginalised groups by AI will allow such discrimination to continue

to grow and by denying the equal treatment of people due to their race, gender, or age will simply allow the problem of inequality to grow bigger. As we have already outlined before, AI is a huge part of our modern lives with systems today which are used to evaluate credit scores and access to loans. If such technologies create results based on race and gender rather than the important figures, then this can cause inequality and a lack of trust within a society. Inaccurate decisions in regards to AI can come in the form of medical treatment or the sentencing of criminals. AI is used in the medical world to diagnose patients based on the symptoms they are experiencing and inaccurate decisions can lead to the need for unnecessary treatment and even death. Furthermore, within the justice system, AI can be used to assess the risks of individuals when determining the length of their sentences. Bias algorithms can result in longer sentences for individuals. These are both examples of human rights violations that occur directly from problems relating to AI. Loss of trust in AI means that people will be less inclined to use the services that AI have to offer due to poor experiences they have had in the past. Using the hospital symptom diagnosis as an example, people may delay researching their symptoms and opt to see a doctor in person instead, which could accelerate the symptoms and lead to further health issues. Bias within AI therefore can create unfair outcomes for different groups of people while also allow social injustice to persist. Hence it is the role of those creating such

systems to make sure that bias is kept to a minimum in order to mitigate the risks that can be created. In his article on whether ChatGPT is an ideology machine, meaning it controls the way we think about things, Weatherby brings up the point that “GPT systems, because they automate a function very close to our felt sense of what it means to be human at all, may produce shifts in the very way we think about things” (Weatherby, 2023). If systems like ChatGPT can impact the way we think about things, then bias within these systems is something to be feared as the spread of incorrect and discriminatory ways of thinking may occur.

Before we can look at how different components of AI can contribute to generating bias outputs, we need to have a case study which we can compare these to. Machine vision is the focus on AI having the ability to understand and interpret visual information such as photography and film. With the development of such technology, systems are able to identify objects in pictures, learn to track objects and understand what is happening in a photo or video. Machine vision is achieved by exposing AI algorithms to datasets of images, film, etc. which are labelled so that the AI can learn to recognise different objects. Machine vision is taught using a combination of the two types of AI which Rossi spoke about – machine learning and symbolic AI. This

is evident, as machine vision utilises large datasets so that the AI can learn to identify different objects, however, the way in which the images are labelled so that the algorithms can categorise each image is representative of symbolic AI. Malevé makes the comparison between the iris and a camera lens, “if the iris can be considered a camera lens, then the world can be grasped as a collection of photographs” (Malevé, 2023). This suggests that the way in which we view life through our eyes is similar to a collection of photographs. Humans learn to understand what different objects are by seeing through their own eyes and being told what everything is, the same way the machine vision learns to identify objects through their own ‘eyes’ – large datasets of visual information.

Data is the first area that can lead to bias that I will be looking at. In regards to machine vision, data correlates to the individual visual information that comes together to form the large datasets; images, film, etc. There are three main components that can cause data bias and they are under or over representation, incorrect labels, and aged data. Under or over representation within data occurs when gathering information to teach AI. For example, if within the training data a machine vision algorithm is shown photos which show one predominant demographic group, it may misinterpret or miscategorise people of other demographics. With image recognition, if an

algorithm is shown images of police which mostly show men as police, the AI will learn to think that the idea of police is associated with men as opposed to men and women. Incorrect labels are the fault of human intervention when labelling images to teach algorithms what are in photos. This can come in many different forms and this will alter how the AI learns to view certain images, causing it to generate false outputs. Aged data refers to the idea that AI are taught using information that is not historically correct any more. This could be the idea that women are no longer seen as lesser than men but instead equal. Karen Hao – a senior AI editor at MIT Technology Review – talks about the idea of "more is more" that LMs are incorporating, "more data to produce bigger models to produce better results". This concept is done by "averaging data from entire populations which has side-lined minority and marginalized communities even as they are disproportionately subjected to the technology's impacts", with experts arguing that "these impacts are repeating the patterns of colonial history" (Hao, 2022). The methods used to collect this data is already "side-lining" marginalized communities and the growth of the data used to teach algorithms can create more bias if such groups are underrepresented.

Similar to data, dataset bias refers to the data as a whole rather than individually. The causes of dataset bias are more or less the same as data bias

as data makes up the datasets therefore dataset bias stems from the way that data is collected and labelled. The difference however is that dataset bias is on a much larger scale. If datasets are full of data which over represents incorrect ideologies, then AI models will be bias. For example, GPT-2's training data was sourced from outbound Reddit links, this data will have a large amount of data which is sexist, racist, etc. Within machine vision, if a dataset is filled with data that incorrectly represents groups it will learn to generate outputs that reflect such ideas. This is bias. AI doesn't generate its own ideas but regurgitates information that it has already been taught, "AI does not invent language ex nihilo but relies on pre-existing databases that are dominated by inscriptions of colonialism, racism, and capitalism" (Plaue, 2021). This is the danger that datasets bias poses as it gives AI the ability to unintentionally spread misinformation.

Model bias can come in various forms however I will specifically be looking at algorithmic bias that can occur. Algorithmic bias occurs when the algorithms used in an AI model favour certain groups over others, be it due to race, gender, etc. In regards to machine vision, if an AI model is designed recognise human faces, mobile phone facial recognition for example, but is taught using a dataset with predominantly white faces, then it may struggle to identify faces of other people. This is one example of the ways in which model bias can occur

and the way that different groups are treated is not fairly done. “The ability of GPT-2 to produce texts indistinguishable from those composed in conventional ways was initially considered so dangerous that OpenAI withheld the model’s public release, for fear that its outputs would trick people into believing that they were real.” (Parrish, 2021), GPT’s ability to generate human-like text is scary as it can replicate biases that it has learnt through its algorithm.

The likes of data bias, dataset bias and model bias can contribute significantly to the creation of bias outputs. These bias predominantly stem from the data that builds up the datasets and the datasets themselves which are used to teach AI. In order to ensure that AI is ethically correct and without bias, supervision during the development and teaching phase is crucial to ensuring that an AI model is fair.

Bibliography:

- Bender, E., et al, 2021, *On the dangers of Stochastic Parrots: Can Language Models Be Too Big?,* Association for Computing Machinery, pp. 611
- Hao, M., 2022, *A New Vision of AI for the people.* The MIT Review
- Malevé, N., Sluis, K., 2023, *The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory,* Vol. 1 (1-2)
- Parrish, A., 2021, *Language Models Can Only Write Poetry.* Available at: https://posts.decontextualize.com/language-models-poetry/
- Plaue, E., Morgan, W., 2021, *Secrets and Machines: A Conversation with GPT-3.* E-flux journal, Issue #123

- Rossi, F., 2019, *Building Trust in Artificial Intelligence*, Journal of International Affairs, Vol. 72(1) pp.127
- Weatherby, L., 2023, *ChatGPT Is an Ideology Machine*. Available at: https://jacobin.com/2023/04/chatgpt-ai-language-models-ideology-media-production
- Yeung, D., et al., 2021, *Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications,* RAND Corporation