CS340 Computational Ethics Assignment 1

Name: 钟志源 Zhiyuan Zhong

SID: 12110517

Multiple Choice Questions

1. C
2. D
3. AB (Virtue ethics is currently one of three major approaches in normative ethics [from Stanford Encyclopedia of Philosophy], I am not sure if C should be inclueded.)
4. D
5. B
6. D
7. ABC (for C: human input could be biased, for D: law can shape how data is managed and used, which can indirectly caused bias, but not directly.)
8. B
9. C
10. ABCD
11. A
12. B
13. B

Short Answer Questions

14.

Generative AI comsumes huge amount of data during training, so I think the bias in these models is mainly caused by the bias in the training data. The training data is collected from the real world, and the real world itself is biased. For example, an experiment on images generated by Stable Diffusion revealed that men with lighter skin tones represented the majority of subjects in every high-paying job, including “politician,” “lawyer," “judge” and “CEO"[1], which maybe because of more training corpus relates white people to these high-paying jobs. Also, certain groups or categories of people may be underrepresented or overrepresented in the dataset. Anyway, this bias is fundementally caused by the bias in the real world(humans).

The bias in the training data is then amplified by the model during training (algorithm), probably because the model is trying to minimize the loss function using gradient descent. The model will learn to generate images that are more similar to the training data, and thus the bias will be learned by the model and reflected in the output.

For the Gemini case, it seems to be an issue of political correctness. "Black Lives Matter" campaign has been a hot topic in recent years, and Gemini might be intentionally prompted internally (system prompt) to generate images of people of all colors (especially people of color) to avoid being accused of racism. This is a kind of bias maybe caused by the policy of Google. Or maybe Gemini is overly aligned with Google's policy by engineers, which is also a kind of bias.

In summary, training data, algorithm, and system prompt(designed by users) are the main sources of bias in generative AI.

References: [1] https://www.bloomberg.com/graphics/2023-generative-ai-bias/