Collaborative Learning Discussion 1

Codes of Ethics and Professional Conduct
Case Study: Malicious Inputs to Content Filters

My Peer response for a student on their topic: ***Malicious Inputs to Content Filters***

Thank you for your in depth review on the topic on malicious inputs to content filters. You raised many good points, and two of those points had focused on, what could potentially be, very worrying areas of ethical and professional behavior. Those points are, the discrimination against protected classes and the securing of the machine learning model.

You had stated that “BP should have reverted to the older, analog model of filtering until the ML model could be reinforced against outside abuse”, which would align with 1(a), 1(b) and 1(c) of the public interest code of conduct (BCS, 2022) and 2(d) of the professional competence and integrity (BCS, 2022). A question arises around the potential of machine learning poisoning of their data model and blacklisting. According to Goldblum, et al. (2023), data can be manipulated to control and degrade the downstream behaviors of learned models, which seemed to be part of the action carried out by the activist. The question therefore is, how long may it take to reverse that data modeling poisoning?

You also stated that “though BP does have the requirement to safeguard children’s access to inappropriate online content, this cannot override potential discrimination against protected classes and politically-charged content.” which would align with 2(d), 2(e) of the professional competence and integrity (BCS, 2022). however, it may be possible that unless a documented policy or standard is written into a service level agreement or law, then it may be argued that the phenomenon known as ‘framing’, could also play a part in deciding what type of content is believed to be acceptable from an individual's stand point (Vanclay et al, 2013).

References

BCS. (2022) Code of Conduct for BCS Members. Available at: https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf [Accessed on 24th June 2023]

Goldblum, et al. “Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses.” *IEEE transactions on pattern analysis and machine intelligence* 45.2 (2023): 1–1. Web. Available from: https://ieeexplore-ieee-org.uniessexlib.idm.oclc.org/document/9743317 [Accessed on 1st july 2023].

Vanclay, F., Baines, J. & Taylor C. (2013) Principles for ethical research involving humans: ethical professional practice in impact assessment Part I. *Impact Assessment and Project Appraisal* 31(4): 243-253. Available from: https://www.tandfonline.com/doi/full/10.1080/14615517.2013.850307 [Accessed on 21st June 2023]