Task 1
I watched it

Task 2

a) **A simplistic approach to moderation is to allow a free-for-all, with no moderation of content whatsoever. Identify a general human right that would potentially undermine, and explain why. *(A general human right shouldn't be specific to Canada or any other place.)***

A general human right that would potentially weaken in a simplistic approach to moderation would be the human's right for no discrimination. For a social media platform with no moderation of content, it is inevitable for users or trolls to start discriminating other users. Posting, commenting, or sharing words or images specifically targeting a specific race, color, sex, language, religion and more.

b) **Suppose that a rival platform, GreaterGood, has a policy where all content is moderated using a purely utilitarian approach:**

If a user-post has zero or positive utility it can remain, otherwise it will be removed. **Assume that the utility of a user-post (including the utility of its wider impact) can be measured. What is the problem with GreaterGood's approach?**

The problem with GreaterGood's approach is that only positive and useful posts would be allowed to stay on their platform. There would be very limited users, some people might even post harmless text but after the measurement of utility would still result in a negative utility and be removed. The freedom of speech here is very limited.

c) **Twitter, Facebook, Instagram, TikTok, WhatsApp,... they're all global platforms, with users from all over the world. What difficulties does that global aspect present when setting rules for which things can be said or done, and which things cannot be said or done?**

The global aspect of these platforms makes it hard to set specific rules because some words that can be said in other countries might not be a good idea to be said in another. Each country in the world has their own culture and ethics, what might be a joke in one might be offensive in another. Although this is no different within a country, some people take offense in certain comments but others see and treat them as a joke. In a different case, some social media platforms are banned in certain countries, like how instagram is banned in China and how several communication platforms are banned in UAE and KSA.

Task 3

Suppose that Twitter decides that some things (certain words or ideas) cannot be said on its platform and will be removed.

With specific reference to the scenario described in the Jon Ronson video, what is the key difficulty with each of these approaches:

a) Report System, where a user reports content after it's been posted for a human to check.

The main difficulty for the in an economical perspective is that in a case where there are millions of tweets everyday and if users report millions of tweets everyday too, it just doesn't make sense that another person would be able to check every single report. Hiring more people just for checking reports is not good for the company either since it isn't everyday that people would report. Also, the decision for the report will not be very accurate, what if in the eyes of that person the post reported is alright but not to the person that reported it. The reporting system also does not make sure that all malicious or negative posts are removed since it will only be brought to twitter's attention once it was reported. In this case the person that said they hope that she is raped was not even reported even though it's literally like committing a crime by telling someone to rape her.

b) Auto-Detection, where an algorithm decides what is acceptable to post.

With auto-detection the one who decides if a post is inappropriate or not is a machine, and I don't think machines are the best thing at decision making. A post that has known negative words like b*tch, or rape would be filtered by auto-detection but not sentences like"I hope you get fired", "good luck with your job hunt", or any comments that has no "bad" words but have a greater negative effect on the target.

Task 4

Which approach, either from the options in question 3 above, or an alternative of your choice, do you think is best for Twitter to take *in general* (not just for the video scenario), and why? Would moving towards a decentralized structure, such as Mastodon help? Or would that make things worse?

The best approach to this would be using both options in question 3, using auto-detection before a post is done and also when a post is reported to filter the posts with negative or "bad" words in it. Also, hiring a human to check posts that are reported but the auto-detection can't detect any inappropriate sentences or words of. Having a decentralized structure could help more than cause harm. Decentralization could help in a way that the people who have the same interest could stay more together and have less interaction with people with different interests

and avoid more negative comments. Also, even if there will be more shaming comments they would stay in their own server and can be filtered by users by not going into that server therefore minimizing the space that the negativity is being spread on.