**Exercise 1.2**

**Part 1**

1. **Exposure of Personal Information**  
   When machine learning is used to moderate online content, it often processes a huge amount of user data—text, images, videos. While the focus is on removing harmful or inappropriate content, there’s a chance that personal information might get exposed, especially if the data isn’t properly anonymized. This could include location details or private messages. To stay on the safe side, companies need to clean and anonymize the data before using it to train their models, and make sure strong privacy protections are in place.
2. **Regional or Cultural Biases**  
   Machine learning models often reflect the data they’re trained on. If most of that data comes from just a few regions or cultures, the model may develop a narrow view. For example, something completely normal in one culture could be wrongly flagged as inappropriate by the model. To avoid this, it’s important that the training data comes from a variety of communities and languages, so the model can better understand different social norms and avoid unfair targeting.
3. **Human Bias in Data Collection**  
   Since humans are usually the ones labeling content to teach these models what’s acceptable and what’s not, their own opinions can sneak in. These personal views can shape the way the model learns, possibly reinforcing certain stereotypes or unfair judgments. One way to reduce this risk is by having diverse teams label the data and regularly reviewing their decisions to make sure they’re consistent and fair.
4. **Potential for Misleading or Harmful Decisions**  
   Machine learning isn’t perfect—it can make mistakes, like deleting helpful content or letting something harmful stay up. If these systems run without any human oversight, the consequences could be serious, especially for vulnerable users. That’s why it’s important to pair machine learning tools with human moderators who can review edge cases and adjust the system when needed.