Ethics Assignment

Authors: Sidner Magnéli, Lukas Martinsson

Hours: 12 h each

Part 1: Privacy and history and limitations of AI

All answers are marked in green.

1. Statistical disclosure limitation (SDL) is a collection of methods designed to protect the privacy of individuals whose data has been used in analyses that will be released to the public. Which of the following specifically compromise the usefulness of SDL to protect? the privacy of individuals

- Increasing computational power
- A critical bug in broadly used SDL software’s
- Fast internet connectivity to mobile devices including LTE and 5G
- Availability of personal data on the internet and other places
- Increased public concern about privacy following revelations by Edward Snowden

2. Which of the following are ethical arguments for protecting subjects' private data in commercial and research applications?

- If subjects do not want their data to be disclosed publicly, they should not share it.
- Protecting the privacy of subjects is a matter of safeguarding their dignity and welfare.
- Private data should only be protected insofar that it does not interfere with a company's ability to maximize economic growth, because human well-being is associated with increasing economic prosperity.
- Protecting private data is only important when it concerns highly private information such as sexual orientation, religion, or political beliefs.
- A failure to protect the privacy of subjects could render them vulnerable to identity theft and extortion.

3. The following analysis techniques or data acquisition techniques are not differentially private

- In a movie rating dataset, removing the names of subjects and randomizing some of their ratings.
- In personal health records including hospital admission, removing names, blood type, annual income, information about HIV status, and social security numbers of subjects.
- In a financial transaction dataset, removing the name and address of subjects excluding postal code, and randomly exchanging labels of credit card transactions within subjects transactions, such that individual payment amounts cannot be tracked to payment of particular services or goods.

- In an online survey measuring adolescents and young adults recreational use of illicit drugs using a yes or no question and registering their age. The subjects' answers to the yes or no question are stored with a probability of 50%. If a subject's answer is not stored it is sampled randomly with a 50-50 chance of yes or no. Similarly, the age is either registered as their correct age with a probability of 50%, else a fake age is sampled from a uniform distribution of integers from 14 to 29 years.
- A series of street-level cameras in down-town Gothenburg are set up to disincentivize crime and simultaneously to collect data to reduce crowding and improve citizens shopping experiences. The cameras use a facial recognition algorithm to track pedestrians faces and count how often subjects pass by a given camera. No images are stored, each camera generates a coded representation of each face using the same algorithm on each camera. The coded representation of a face, along with the GPS coordinates of the camera, date, and time are stored.

4. Fisher made important contributions to statistical significance testing. He saw it as a way of communicating 'statistical findings that was as unassailable as a logical proof'. Statistical significance testing can not:

- Tell us whether a sample is likely to have occurred under the null hypothesis.
- Tell us whether what we are measuring is important with regards to the conclusion we wish to draw.
- Tell us whether a sample is unlikely, up to some probability threshold, to have occurred under the null hypothesis.
- Infer the underlying cause of the measured data.
- Objectively relate human's ethnicity's relationship to their personality traits.

5. What were important early drivers of the statistical research of Galton, Fisher, and Pearson?

- The study of coin-flips
- Eugenics
- Association of genetic variants observed in large human cohorts to disease phenotypes from next-generation sequencing technologies.
- Biological evolution.
- Fault detection of machines in factories powering the industrial revolution.

6. In a study from 1925 published in the scientific journal which is now known as "Annals of Human Genetics", Pearson claimed to have shown that Jewish children, in particular girls, were less intelligent on average than their non-Jewish counterparts. He further claimed that intelligence was not significantly correlated with any environmental factor that could be improved, such as health, cleanliness, or nutrition. What aspects of the study may influence these conclusions?

- Pearson did not use the right statistical significance test.
- Pearson did not use a certified intelligence test.
- Pearson made use of self-reported data (survey data) of the conditions of student's home lives.
- Pearson made use of teachers' assessment of student's intelligence.

- Pearson did not account for environmental factors that could not be improved, e.g., the students' refugee status.

7. In the Fragile Families Challenge researchers from 160 teams were tasked to predict a number of social outcomes -- child grade point average, child grit, household eviction, household material hardship, primary caregiver layoff, and primary caregiver participation in job training -- using extensive longitudinal data. Is this a difficult prediction task? Why? Why not?

- Yes, because the challenge includes predicting the future (“wave 6”), and only selected data is available for the time-point (“wave 6”) to be predicted.
- Yes, because the challenge includes predicting the future ((“wave 6”), and no data is available for the time-point ((“wave 6”) to be predicted.
- Yes, because social outcomes depend on many parameters. Many of which may not be captured well by the dataset.
- No, because the researchers can use any machine learning method and they have access to a big data set.
- No, because social outcomes are known to be easy to predict even with simple models.

8. The authors of the Fragile Families Challenge study argue that 'predictability' is poor measure for understanding the mechanisms of social outcomes. What alternatives do the authors propose?

- Descriptions.
- Using statistical significance testing.
- Using Foucault's analysis of power.
- The current understanding is correct but incomplete because it lacks theories that explain why outcomes are difficult to predict even with high-quality data.
- Causal inference.

9. What important ethical implications may a poor 'predictability' of social outcomes have?

- It will be more difficult for employers to use AI to screen job applicants, leading to an increased workload on Human Resource employees.
- Using predictive modeling in the criminal legal system may lead to wrongful convictions and arrests.
- It will put legal strain on the police and military because the most performant models for identifying criminals and terrorists use deep learning that are difficult to interpret.
- Using predictive modeling in the child-protective services may lead to wrongful removal of children from their families.
- It will require larger machine learning models to improve predictions, which in turn require more compute power and electricity to run, putting strain on the environment.

10. What reasons can lead to unfair and unethical AI and Data Science models and algorithms?

- Sampling bias.
- Systemic bias in society.
- Poorly designed surveys.
- Low memory capacity of embedded devices.

- Overly rigorous testing and use of the common task method.

Part 2: Wearable health-technologies and private health insurance

1) ***How can the training pipeline affect the predictions of Vivaksa's health status prediction algorithm when deployed on the general public? (1.5 pts)***

The decision to let the company's employees, a group likely to be somewhat biased towards healthy lifestyle, generate the training data might contribute to a poor and biased model. Firstly, one very important factor for good model performance is the size (number of employees) of the training data. Secondly the firm's employees are hardly representative of the true population, as health markers vary between parameters such as age, employment and even socio-economic background. Another bias, due to the short time span of the training data (one month), overvaluation of temporary deviations from normal daily activity among the employees. As a consequence of this their model might make inaccurate predictions when used on the general public.

2) **Can Vivaksa improve their predictions? if so, how? and if not, why? (1 pt)**

Two possible improvements that might counteract the previously mentioned problems are increased time span and improved sampling of the training data. For the above scenario, increased time span means not only using the last month of activity data for training, but rather increase that span to months, or even years. This might, in reality, not be possible due to limitations in time or existence of historical data. Improved sampling refers to sampling training data from a population that better represents the true population (the general public or the expected consumer group). To find out the expected consumer group a market analysis could be performed. That being said, it is difficult to find individuals so that the dataset truly represents the whole consumer group, which in itself might change with time. Important to note is that it does not have to be precisely representable since this could interfere with unbiased selections of individuals, although as close as possible is preferable.

3) **Can the roll-out of this pilot program undermine the insurance company's own value of fairness? - explain why or why not (1 pt)**.

It depends partly on how accurate Vivaksa's model is as a predictor of health for the individuals that use it. Since the training dataset is probably not representative of the individuals that the model will be predicting on, it's quite likely that it **can** affect the insurance company's own value of fairness. Secondly even assuming that Vivaksa is accurate it is only able to check certain vital signs. Therefore, some individuals with worse health could get better premiums than some individuals with "better" health, since they might show better vital signs of those that Vivaksa can detect. This undermines fairness. Lastly there is the discussion if it is fair to change someone's premium the second they start feeling worse. This could even incentives previous healthy people whose health take a turn for the worse to opt-out, even if it means paying a higher premium (since that premium could be lower than if Vivaksa could find out their current health status)

Part 3: AI-enabled human personality prediction

1) **What are the limitations of Selfie2Personalitys technology - if any? (1pt)**

   Since no meta-analysis of machine learning about predicting trustworthiness and criminality through facial features showing its usefulness and accuracy has been done, the inspiration and Selfie2Personalitys' own machine learning could likely give wrongful predictions masked as accurate ones. Furthermore, it is well known that individuals mainly post "good" pictures of themselves on social media. This suggests a drawback, the implicit homogeneity of the pictures accessible for prediction. Also most likely, the app will improve through the data gathered from the users. However, the userbase would most likely not be representative of the population, making predictions of less represented groups (when it comes to age, gender etc) worse. Lastly, since the app indicates that it will provide information to law enforcement instead of allowing them to use the AI system it is less likely that felons would want to use the app, which in turn would make it worse at predicting them.

2) **Selfie2Personality share the social media accounts and a psychological profile of users from the city metropolitan area who fist the political stance (as predicted from their app) that law-enforcement found to be sympathetic to the protests. Are there any ethical and privacy concerns with this collaboration and how may they manifest? (2 pts)**

   One aspect that should raise ethical concern is that individuals could be incorrectly classified as "sympathetic" due to reasons outside their control such as insufficient data leading to low accuracy predictions or model biases. For example, confirmation bias, whether it arises from a skewed training dataset or unconscious wrongful analysis by law enforcement. As a result, incorrectly classified "sympathetic" individuals could start acting differently, which would/could confirm the app's assessment, when in reality is a response to its initial assessment. Furthermore, the app would more likely think friends and family of a "sympathetic" individual are also sympathetic, compromising not only the individual, but also the privacy of the individuals' network. If instead of a political stance this would be a question of ethnicity, then this would immediately be considered unacceptable, which is an indicator of the severity of the collaboration.