Jude Wells Programming with Data – Coursework I Data Ethics Assignment

Hooa-Mai (HM) is a website offering personality analysis. Users answer a detailed questionnaire and in return the site gives them information/insight about their personality.

The analysis is based on sophisticated use of aggregated data from the users (possibly combined with other publicly available data). This requires a significant technical team, the cost of which is largely covered by the resale of anonymized user data. The site is seeking to position itself as a serious player with which users can monitor the evolution of their personalities over time, thereby standing out in what is often regarded as a frivolous sector.

The users are aware that their data will be resold and they understand that, as with all resale of data, there is a small but real risk that, in combination with other publicly available data, this might lead to their identification or to other disadvantages for them. They accept this in principle, but may well be more concerned about some possible risks than others.

Now think about two positions that you could be in with regard to HM: (a) one of the founders and key drivers, with particular responsibility for data ethics and (b) a long-term user. You are asked below a number of ethics-related questions.

Please answer each question in no more than 200 words. The questions are designed to bring out moral intuitions, but you should aim not merely to state your intuitions but also to give reasons for them and also consider possible objections/differing perceptions. Marks will be awarded on the basis of the extent to which you are able to outline a principled approach to deciding what is right or wrong in each particular case.

Position A: You are a founder of HM with ethics responsibilities

1. To what extent do you have a positive duty proactively to explore areas where the combination of user data with other data could lead to problems?

The broad principle of data ethics is that the data-subject (in this case the HM user) has rights and protections relating to how their personal data is used and shared. A company such as HM cannot do as it pleases with user data if that data relates to an identifiable individual. Many companies such as HM may seek to annonymise and aggregate user data so that it it is not subject to the tight legal restrictions relating to personal data. This may allow them to share the data with third parties or use it for commercial purposes. However, it is the company's legal and moral duty to ensure that annonymised aggregated data cannot be used to identify an individual. Removing a person's name is not sufficient: there are multiple ways that a data-subject might be identifiable, for example: phone number, national insurance number or IP address. A further risk comes through the combination of multiple generic pieces of data. For example, it would not be hard to identify this hypothetical data subject 'female, aged 18-24, lives in postcode area WC1E, has a disabled persons parking pass, drives a red Ford focus.' Another potential risk arises where annonymised data can be combined with data from external sources to identify a subject: for example an IP address could be checked against an ISP's list of customers to find the name and address of the subject.

2. To what extent and in what circumstances should you inform the users when you consider that there are new possible risks arising from the use of their data?

A company has a legal and moral duty to conduct risk assessments that consider the possibilities of things such as data breaches, de-annonymisation or exploitation of users. In many (but not all cases) EU law requires companies to obtain informed consent from users before personal data can be stored or processed. In such cases, the consent would no longer be valid if the nature of the data-processing has significantly changed or if new risks have arisen which the user was not aware of when they initial gave consent. In these cases the key question is: have the risks changed to such an extent that the user's previous consent can no longer be said to cover the current situation?

3. To what extent should you differentiate between the interest in confidentiality of different categories of users?

For certain users, breaches of confidentiality may be particularly damaging and therefore additional protections would need to be in place. The EU General Data Protection Regulations recognizes a class of 'sensitive personal data' which requires additional protection and is more strictly regulated. This includes information regarding race / ethnic origin, political opinions, sexual life and health.

It should also be noted that some users may not be able to give informed consent: for example children or adults with cognitive impairments. It may be the case that these users cannot meaningfully consent to having their data passed on to third parties.

In some cases it is appropriate to breach confidentiality in order to protect someone from harm. This principle is recognized in UK law which deals with safeguarding of children and vulnerable adults.

4. If you discover that a given user fits into a pattern associated with other users who encounter a particular medical problem, should you proactively inform the user of this risk?

The potential negative consequences of informing users that they may have a medical problem are that it is likely to cause distress, it might have an effect on a user's private medical insurance and there could be other serious consequences if the user makes decisions on the basis of an incorrect diagnosis. Many would argue that a company that is not involved in the medical field has no right to make forays into this area: they might point out that if a person goes to see a doctor they have consented to being examined and diagnosed, this is clearly not the case with someone who is simply a user of HM. On the other hand, we must consider the consequences of not informing a user: inaction is a decision and we are not absolved of responsibility because we did nothing. Informing a user may give them the opportunity to seek treatment. A good approach would be to obtain informed consent from a user, before any of the data is processed. The user could opt-in to being notified of medical issues and hopefully this would ameliorate any feeling of intrusion. If a user does not consent to being notified then it is probably best to avoid processing their data in any way that could lead to this situation.

Even with informed consent, the company has to make a calculation as to what level of statistical significance passes the threshold of informing the user. This requires a careful cost-benefit analysis which considers the likelihood of false positives and the cost they have to the user.

5. Would you feel it appropriate to use data provided by the users in support of a political or other cause of which you strongly approve?

In answering this question there are multiple factors that would cause each case to sit differently on the ethical spectrum. It will be helpful to examine each of these in turn to build a framework that will allow us to assess whether the use of data is appropriate:

1) To what extent is the cause controversial? Most people would recognize the distinction between research (where the purpose is to uncover truth, hopefully with some degree of objectivity) and campaigning (where information is used to advance a particular position). Using information to campaign is intrinsically biased, however there will be some situations where a cause may be less controversial, for example a charity that uses data to advocate for improvements in education. Other causes such as campaigning for a political candidate will be more controversial; therefore utilizing data for this cause would be less appropriate.
2) How personal is the data? – it's a reasonable generalisation to say that the more personal and private the data is, the more unethical it is to use it for a cause. An extreme example of unethical use of data would be leaking private medical data of a political opponent to harm his or her campaign.
3) To what extent did the data subject consent to this use? If the subject gave fully informed consent this can permit the data to be used for a wide range of causes. Fully informed consent means that the subject understands exactly how the data will be used and the subject has given explicit approval. In cases where consent is given by a less informed subject, or the consent is given in broad or vague terms then the use of the data should be more constrained.

Position B: You are a user of HM. The following possible uses of your data all involve a slightly increased risk of deanonymisation.

1. Would you be happy for your data to be used for strictly medical research?

In most cases, I would be happy for my data to be used for medical purposes because I'm supportive of the endeavor of advancing medical science. Reaching this conclusion involves weighing up the benefits (advancing medical science) against the risks. I would consider that there is relatively little risk to me personally if there was a breach of my medical data. As an individual I also believe that medical research (compared with social research and commercial research) is more regulated with appropriate oversight and ethical considerations.

Other people may be more cautious about allowing their data to be used for medical research. Some people have religious, moral or cultural objections to fields within medicine. Stem cell research, euthanasia and genetics are three examples where people often have objections. Others may have particular reasons where the harm from a data breach or identification are so high that it does not justify the risk. If a subject's personal medical information contained sensitive information such as a diagnosis of a communicable disease they may be more concerned with ensuring that they are not identifiable in the data.

2. Would you be happy for your data to be used for more general social research?

If my data were to be used for social research I would like to know the nature of the data being shared, who would be using it, and for what purpose. I would be more cautious about sharing data if it was not aggregated and anonymous. I would feel somewhat aggrieved if I were to find out that my personal data was being used for a purpose which I had not consented to. However, I also am mindful of the fact that increased legislative regulation of data slows down academic progress and can cause an administrative burden on organisations. I feel that research is a positive endeavor and from my experience of being involved in academia as a student, I understand that research is often held back by a lack of available data. The internet opens up many opportunities to cheaply gather data but researchers should be cautious not to exploit this in a way which might foster distrust and calls for tighter regulation.

3. Would you be happy for your data to be used for commercial purposes, which might lead to your receiving advertisements/offers for products/services tailored to your predicted preferences?

We have probably all experienced the phenomena of being eerily followed around the internet by an advert for an item that we once viewed on the web. Many people feel that this is an intrusion as it leaves a feeling as though one is being monitored and tracked for the purpose of selling more products. At the same time, there is also a degree of convenience in being presented with relevant products, songs or films that are algorithmically determined to be inline with our tastes and needs. The downside to 'predicted preferences' is that we may find ourselves in an echo chamber where existing preferences are reinforced while opportunities for new discoveries are diminished. Occasionally, data gathered for commercial purposes can stray into the most personal realms of private life, such as the infamous (possibly apocryphal) story about the supermarket that predicted a customer's pregnancy and sent her maternity related vouchers before she had even told her family. As with each of these cases, consent is key, particularly in cases where personal data is being shared with a third party.

4. Would you be happy for your data to be used for charitable purposes in line with your perceived sympathies?

In general I can see the benefit of allowing charities access to more data. Of course 'charitable purposes' could include being targeted by fundraisers working on behalf of the charity, many people find this to be an intrusion particularly because emotive arguments are often used by fundraisers to solicit donations. Charities should be cautious not to abuse the trust that people hold in them. It's possible to envisage unethical charity fundraising where personal information (such as knowledge of a bereavement) could be used to target and emotionally manipulate people into donating money. I would have further concerns over the prospect of a company deciding what are my 'perceived sympathies' given that this implies an element of interpretation by the company.

5. Would you be happy for your data to be used in the interest of political parties which you support?

Data in political campaigning is an issue which has become increasingly prominent and controversial in recent years. It can be used as a tool for democratization for example in cases where political parties develop data-driven policies and manifestos; either through social research or using data to understand sentiment and preferences of the public. On the other hand, there are also examples where candidates present different political messages (either online or on the doorstep) because they have data which indicates the political preference of the targeted would-be supporter. I think that ethical principles should guide an individual's assessment of whether political use of data is right or wrong, the logical corollary of this idea is that if a particular use of data is unprincipled, or has a detrimental impact on democracy then it should be opposed whether it is being used by a party that you support or oppose.

Submission: please upload your essay on moodle, in PDF formatting if possible. Format your answer sheet to make it easier to connect your answers to the relative questions.

Plagiarism: please be advised that moodle deploys state-of-the-art plagiarism detection software to evaluate coursework submissions, both against Web sources and against other submissions, past and present. Each submission will be scored for originality; submissions with low originality might be discarded.

It is however possible to insert quotations by using appropriate typographic style and providing the reference:

this phrase is an example of a typographic style for citation and reference [Lawson-Tancred & Provetti, Coursework I, 2018].