CSAI - Tutorial 1 (06 Feb 2025)

We will be analyzing a short version of OkCupid case study introduced in Part Three of An Introduction to Data Ethics book by Prof Shannon Vallor.

OkCupid

In 2016, two Danish social science researchers used data scraping software developed by a third collaborator to amass and analyze a trove of public user data from approximately 68,000 user profiles on the online dating website OkCupid. The purported aim of the study was to analyze "the relationship of cognitive ability to religious beliefs and political interest/participation" among the users of the site.

However, when the researchers published their study in the open access online journal Open Differential Psychology, they included their entire dataset, without use of any deanonymizing or other privacy-preserving techniques to obscure the sensitive data. Even though the real names and photographs of the site's users were not included in the dataset, the publication of usernames, bios, age, gender, sexual orientation, religion, personality traits, interests, and answers to popular dating survey questions was immediately recognized by other researchers as an acute privacy threat, since this sort of data is easily re-identifiable when combined with other publically available datasets.

That is, the real-world identities of many of the users, even when not reflected in their chosen usernames, could easily be uncovered and relinked to the highly sensitive data in their profiles, using commonly available re-identification techniques. The responses to the survey questions were especially sensitive, since they often included information about users' sexual habits and desires, history of relationship fidelity and drug use, political views, and other extremely personal information. Notably, this information was public only to others logged onto the site as a user who had answered the same survey questions; that is, users expected that the only people who could see their answers would be other users of OkCupid seeking a relationship. The researchers, of course, had logged on to the site and answered the survey questions for an entirely different purpose—to gain access to the answers that thousands of others had given.

When immediately challenged upon release of the data and asked via social media if they had made any efforts to anonymize the dataset prior to publication, the lead study author Emil Kirkegaard responded on Twitter as follows: "No. Data is already public." In follow-up media interviews later, he said: "We thought this was an obvious case of public data scraping so that it would not be a legal problem."[1] When asked if the site had given permission, Kirkegaard replied by tweeting "Don't know, don't ask. :)"[2] A spokesperson for OkCupid, which the researchers had not asked for permission to scrape the site using automated software, later stated that the researchers had violated their Terms of Service and had been sent a take-down notice instructing them to remove the public dataset. The researchers eventually complied, but not before the dataset had already been accessible for two days.

[1] Hackett (2016): http://fortune.com/2016/05/18/okcupid-data-research/
[2] Resnick (2016): https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release

1. State the nature of the ethical issue you have initially spotted
2. List the relevant facts
3. Identify stakeholders
4. Clarify the underlying values
5. Consider consequences
6. Identify relevant rights/duties
7. Reflect on which virtues apply
8. Consider relevant relationships
9. Develop a list of potential responses
10. Use moral imagination to consider each option based on the above considerations
11. Choose the best option
12. Consider what could be done in the future to prevent the problem

Table 1: 12-Step Approach

Discussion Questions

1. What specific, significant harms to members of the public did the researchers' actions risk? List as many types of harm as you can think of.

2. How should those potential harms have been evaluated alongside the prospective benefits of the research claimed by the study's authors? Could the benefits hoped for by the authors have been significant enough to justify the risks of harm you identified above in Question 1?

3. The lead author repeatedly defended the study on the grounds that the data was technically public (since it was made accessible by the data subjects to other OkCupid users). The author's implication here is that no individual OkCupid user could have reasonably objected to their data being viewed by any other individual OkCupid user, so, the authors might argue, how could they reasonably object to what the authors did with it? How would you evaluate that argument? Do you find the data collection method used within this study ethical?

4. A Danish programmer, Oliver Nordbjerg, specifically designed the data scraping software for the study, though he was not a co-author of the study himself. What ethical obligations did he have in the case? Should he have agreed to design a tool for this study? To what extent, if any, does he share in the ethical responsibility for any harms to the public that resulted?

5. Assume that you are the supervisor of the lead researchers in this study, and you need to decide if you should share the study findings with the public. To make such a decision, **apply 12-step approach to this case study** as shown in Table 1.