CSAI - Tutorial 1 (08 Feb 2024)

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

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 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:
   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

Answers

Here we provide some *potential* answers for the discussion questions. Feel free to think about other dimensions about the case study.

1. Three harms that could arise from this case study are:

   - Social: the users of the dating site risk social embarrassment and possible harms to existing relationships. Increased discrimination/prejudice towards members of certain groups because of their religion, political views, sexual orientation or so on.
   - Psychological: publicly exposing personal information, such as those related to user's sexual lives, could lead to anxiety, depression, social exclusion, suicide or so on.
   - Legal: the authors of the study broke OkCupid's Terms of Service and published data to an unintended audience. Both the website and its users may have grounds to pursue legal action. The privacy rights of the users are violated as well, since their data is not anonymous anymore when combined with publicly available datasets.

2. Since the harms affect the users and owners of OkCupid, one would hope that they would in some way benefit from this study. It is difficult to see how this might be the case, perhaps understanding how cognitive ability and religious/political leanings are related may lead to users making more informed choices with regards to finding a partner. However, this study could also lead to increased discrimination and prejudice towards certain groups. In the former case, it still seems difficult to justify the results at the cost of social and psychological well-being. The authors are mostly motivated by their own interests, and they only focus on publishing a paper. They do not consider other stakeholders who could benefit from such a study. The harms identified above are very critical and they should have been considered in the first place.

3. There are a number of problems with the authors' arguments. First, the OkCupid data was only accessible to OkCupid users. It is not public in the sense that anyone can access it since you must first sign up and answer personal questions. The users do this with the assumption that others on the site are also seeking relationships and provide information which aligns with those intentions. This intention is made very clear from the context of the website in their Terms of Service. The way the authors accessed and made use of the data in this case could then be classed as a form of dishonesty and unethical behaviour since they betray the implicit agreement between users that they are all there to seek a relationship. Secondly, they develop a software program to collect data automatically again by violating the website's Terms of Service. The researchers do not get in touch with OkCupid for the proper use of data for their research. Thirdly, they make the users' data publicly available; which was supposed to be only accessed by OkCupid users.

4. Nordbjerg had the obligation to ask the authors what they intended to use his software for and then analyse the ethical impacts that the work may have. If the data was not published and kept internal, the ethical implications are more ambiguous. However, if Nordbjerg was aware that the data was to be made public he should have refused to participate. Or he should have made it clear to the authors that such a software could have non-negligible ethical considerations. His culpability in this matter depends on several factors, including how much information he had access to and whether he made reasonable attempts to understand the impacts of the broader work. There is yet another ethical problem in this specific case. A researcher is contributing to a research project by providing significant support, and this researcher is even not one

of the listed authors of the paper. This is a disrespectful act against the researcher. Or maybe this researcher was trying to stay anonymous, but the authors made his identity public as well. Who knows?

5. According to the 12-step approach:

   (1) The main concern is the publication of personal/sensitive user data containing very private information with no direct consent.

   (2)
   - The study seeks to "analyze the relationship of cognitive ability to religious beliefs and political interest/participation" among the users of OkCupid.
   - The researchers scraped user data from the site by signing up as regular users.
   - The data which was scraped was published along with the findings with no anonymization.
   - The authors did not feel it was necessary to ask OkCupid for permission to perform the scraping and sharing the dataset publicly.

   (3) The stakeholders are the website, the website users, their friends and relatives, the researchers (including the person who developed the software) and possibly other researchers within the field who would read and benefit from the research. Other possible stakeholders are: the University where the researchers conduct the research, the Journal team (including reviewers, editors) who agreed to publish the paper; the Society overall who has access to the paper and the data.

   (4) The users and website value their data privacy, they also respect the implicit agreement that all users to the website will not misuse the platform. OkCupid also tries to maintain social trust of their users, who want to feel safe while using their service. The researchers seem to value data transparency and research, they wish to make their data public as they believe it could be accessed by anyone anyway.

   (5) Refer to questions 1 and 2

   (6) The website has the duty to maintain its terms of use and privacy policy. The users have the right for their data to be used only for purposes intended by the platform and made clear by the website's policy. The researchers have the duty: (i) to ensure they make every effort to not harm users whose data they provide, (ii) to seek ethical consent before starting any research including human subjects while providing compensation if possible, (iii) to respect OkCupid team, other researchers and website users.

   (7) Some virtues which are important to this case study are honesty (did the researchers use dishonest means to gain access to the data?), respect (do the authors respect the users' reputation and wishes/other researchers/research community?) and fairness (how fair is it to share users' data without using any privacy-preserving techniques?).

   (8) The relationship with the website and the users could be negatively impacted by the possible breach of trust. Such a breach could have a bigger impact on the society who would not prefer using OkCupid (or any similar website) anymore. The researchers have an ongoing relationship with their University. Unethical research may end up in a cut of fundings for these researchers in the future. The readers of the journal can also lose their trust in the journal, since the editorial team chooses to publish papers conducting unethical research.

   (9) It is possible to generate more outcomes than the ones mentioned below:
   - Publish the data as is,
   - anonymize data before publishing,

- seek permission from OkCupid before publishing with anonymization,
- seek ethical permission from the website and users, and the University; while acknowledging the efforts put by other researchers,
- not publishing the data regardless (harms the ethical concept of integrity).
- ...

(10) The analysis has clearly shown that first option is unethical. The second option could possibly solve the possibility of user traceability, yet with modern de-anonymization techniques there is still a risk. The second option also has the issue of breaking the terms of service outlined by the website and could still be considered a breach of trust with the users. The data is still not collected and used ethically. The third option solves several ethical issues but there is still the concern that the users never consented to their data being used in this way when registering for the service. Option four addresses this by contacting the users who's data would be made public if the users give consent in sharing their data with the researchers. On the other hand, the researchers also get Ethics approval from their University to conduct the research; while mentioning the names of the researchers explicitly who contribute to the research. The option of not publishing would obviously solve all the ethical challenges but would go in direct conflict with the aims of the researchers and still not justifying the collection and use of data in an unethical way.

(11) Seek permission from website and users, remove all non-consenting users and don't publish if permission was not given, share data in an anonymized way, acknowledge the researcher who helped in data collection, get ethics approval from the University where the research is conducted, and so on. This is a moderate approach allowing the authors to achieve their aims while also protecting OkCupid users hence minimizing harms.

(12) In the future, this consent should be requested from the outset. It should not be acceptable to access and analyse personal data that has not been explicitly made public by the creators of the data. One should also consider the policies of the platform on which the data has been hosted and request approval if any ambiguity were to arise. OkCupid can also be more careful against automatic collection of users' data from their website; and ban any suspicious activities to prevent malicious users. Universities should provide the training needed for their researchers to make it clear when they should get Ethics approval to conduct research. Academic venues such as conferences, journals should emphasize that non-ethical research should be flagged during the review process.