Bhuvi Ghosh
60009210191

Tutorial-3

Title: Ethical Dilemmas and Privacy Concerns in Data Scraping: The 2016 OkCupid Study Controversy

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."17 When asked if the site had given permission, Kirkegaard replied

by tweeting "Don't know, don't ask. :)"18 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.

Critics of the researchers argued that even if the information had been legally obtained, it was also a flagrant ethical violation of many professional norms of research ethics (including informed consent from data subjects, who never gave permission for their profiles to be used or published by the researchers). Aarhus University, where the lead researcher was a student, distanced itself from the study saying that it was an independent activity of the student and not funded by Aarhus, and that "We are sure that [Kirkegaard] has not learned his methods and ethical standards of research at our university, and he is clearly not representative of the about 38,000 students at AU."
The authors did appear to anticipate that their actions might be ethically controversial. In the draft paper, which was later removed from publication, the authors wrote that "Some may object to the ethics of gathering and releasing this data…However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form."

Question 3.1: What sort of goods can data professionals contribute to the public sphere? (Answer as fully/in as many ways as you are able):

Data professionals have the potential to make numerous valuable contributions to society through their expertise and work:

1. They can analyze large datasets to identify patterns in public health that lead to improved medical treatments and better healthcare outcomes for communities.
2. Data professionals can help governments and urban planners optimize public transportation systems and city infrastructure by analyzing movement patterns and usage data.
3. They can develop predictive models that help communities prepare for and respond to natural disasters, potentially saving lives and protecting property.
4. Through careful analysis of educational data, they can identify effective teaching methods and learning interventions that help improve educational outcomes for students.
5. Data professionals can contribute to environmental protection by analyzing climate data and helping organizations reduce their carbon footprint through data-driven sustainability initiatives.

Question 3.2: What kinds of character traits, qualities, behaviors and/or habits do you think mark the kinds of data professionals who will contribute most to the public good? (Answer as fully/in as many ways as you are able):

The data professionals who make the most positive impact on society tend to exhibit several crucial characteristics and behaviors:

1. They demonstrate unwavering ethical integrity by consistently prioritizing privacy protection and obtaining proper consent when handling sensitive data.
2. Successful data professionals show strong critical thinking skills combined with humility, always questioning their assumptions and being open to alternative interpretations of their findings.
3. They possess a deep sense of social responsibility, carefully considering the potential impacts of their work on various stakeholders and actively working to prevent harm.
4. These professionals maintain transparency in their methods and are committed to clear communication, ensuring their work can be understood and scrutinized by both peers and the public.
5. They exhibit persistent curiosity and a commitment to continuous learning, staying current with both technical developments and evolving ethical standards in their field.

Question 3.3: 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.

The researchers' actions in the OkCupid case created several serious potential harms to users:

1. Users faced potential personal and professional damage through the exposure of intimate details about their sexual preferences, religious beliefs, and political views that could be used against them by employers, colleagues, or family members.
2. The publication of users' dating profiles alongside their survey responses about drug use and relationship history created a risk of blackmail or exploitation by malicious actors.
3. Users could experience significant emotional distress and violation of trust upon discovering their private information was being used for research purposes without their consent or knowledge.
4. The combination of profile data with other publicly available information could lead to stalking or harassment, particularly for vulnerable users such as LGBTQ+ individuals in less accepting communities.
5. The release of sensitive personal data could damage users' future romantic prospects by permanently exposing private information they had only intended to share in the context of dating.

Question 3.4: 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 3.3?

The evaluation of potential harms versus benefits in this case reveals a clear ethical failure by the researchers:

1. The study's claimed benefit of understanding relationships between cognitive ability, religious beliefs, and political participation could have been pursued through proper research channels with informed consent, making the unauthorized data scraping unnecessary and unjustifiable.
2. The potential for serious personal harm to tens of thousands of individuals far outweighed any possible academic insights that might have been gained from the research.
3. The researchers' failure to implement basic privacy protections or anonymization techniques demonstrates they did not make any meaningful attempt to balance benefits against risks.
4. The study's methodology of mass data scraping without consent violated fundamental research ethics principles, undermining any scientific value the findings might have had.
5. The claimed benefits were primarily academic in nature, while the potential harms were direct, personal, and could have life-altering consequences for the individuals whose data was exposed.

Question 3.5: List the various stakeholders involved in the OkCupid case, and for each type of stakeholder you listed, identify what was at stake for them in this episode. Be sure your list is as complete as you can make it, including all possible affected stakeholders.

The OkCupid case involved multiple stakeholders with varying degrees of risk and interest:

1. The individual OkCupid users had their privacy, personal safety, and reputation at stake, as their sensitive personal information was exposed without their consent or knowledge.
2. OkCupid as a company faced potential damage to their reputation, loss of user trust, and legal liability for failing to protect their users' data from unauthorized scraping.
3. The academic community, particularly in social sciences, risked damage to their credibility and public trust due to the researchers' unethical conduct being associated with academic research.
4. The software developer who created the scraping tool faced professional reputation damage and potential legal consequences for his role in enabling the unauthorized data collection.
5. Future researchers conducting legitimate studies might face increased public skepticism and stricter regulations due to the breach of trust caused by this incident.

Question 3.6: The researchers' actions potentially affected tens of thousands of people. Would the members of the public whose data were exposed by the researchers be justified in feeling abused, violated, or otherwise unethically treated by the study's authors, even though they have never had a personal interaction with the authors? If those feelings are justified, does this show that the study's authors had an ethical obligation to those members of the public that they failed to respect?

The violation of users' privacy and trust in this case raises important ethical considerations:

1. The OkCupid users would be entirely justified in feeling violated, as their reasonable expectation of privacy within the dating platform's ecosystem was deliberately breached for purposes they never consented to.
2. The scale of the violation doesn't diminish its personal nature - each of the 68,000 affected users had their individual right to privacy and autonomy disregarded by the researchers.
3. The researchers had a clear ethical obligation to these users simply by virtue of accessing and handling their personal information, regardless of whether direct interaction occurred.
4. The fact that users voluntarily shared this information within OkCupid's platform doesn't negate the researchers' ethical duty to respect the intended context and limitations of that sharing.
5. The researchers' ethical obligations extended beyond just data handling to include protecting human subjects in research, a fundamental principle they failed to honor.

Question 3.7: 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? Does it make an ethical difference that the authors accessed the data in a very different way, to a far greater extent, with highly specialized tools, and for a very different purpose than an 'ordinary' OkCupid user?

The lead author's defense about public data accessibility reveals several significant ethical flaws:

1. There is a fundamental difference between individual users accessing profile data for dating purposes and researchers mass-harvesting data for academic study without consent.
2. The specialized scraping tools and automated collection methods used by the researchers represent a technological exploitation that goes far beyond the intended and reasonable use of the platform.
3. The context in which users shared their data matters significantly - they consented to share intimate details with potential dates, not to have their data analyzed and published in academic research.

4. The researchers' argument ignores the principle of "practical obscurity" where information that is technically public still retains privacy protection through the practical difficulties of accessing and aggregating it.
5. The mass collection and publication of the data fundamentally changed its nature and potential impact, creating new privacy risks that users could not have reasonably anticipated when sharing their information on the platform.

Question 3.8: The authors clearly did anticipate some criticism of their conduct as unethical, and indeed they received an overwhelming amount of public criticism, quickly and widely. How meaningful is that public criticism? To what extent are big data practitioners answerable to the public for their conduct, or can data practitioners justifiably ignore the public's critical response to what they do? Explain your answer.

The relationship between data practitioners and public accountability is crucial for ethical practice:

1. Public criticism serves as a vital ethical check on data practitioners' work, reflecting society's values and concerns about how personal data is used and protected.
2. Data practitioners cannot justifiably ignore public criticism because their work directly impacts people's lives and privacy, creating an inherent responsibility to consider public concerns.
3. The swift and widespread criticism in this case demonstrated the public's growing awareness of data privacy issues and their right to hold practitioners accountable for misuse of personal information.
4. Professional data practitioners have a responsibility to be transparent about their methods and engage meaningfully with public concerns, as their work relies on public trust and cooperation.
5. Dismissing public criticism risks undermining the entire field of data science by eroding trust and potentially leading to stricter regulations that could limit beneficial research.

Question 3.9: As a follow up to Question 3.7, how meaningful is it that much of the criticism of the researchers' conduct came from a range of well-established data professionals and researchers, including members of professional societies for social science research, the profession to which the study's authors presumably aspired? How should a data practitioner want to be judged by his or her peers or prospective professional colleagues? Should the evaluation of our conduct by our professional peers and colleagues hold special sway over us, and if so, why?

Professional peer criticism carries weight in the field of data science and research:

1. Criticism from established professionals carries particular significance because they understand the technical and ethical complexities involved in data research and can evaluate both the methods and implications of the work.

2. Professional peers serve as guardians of field standards and ethics, helping to maintain the integrity and public trust necessary for the field to continue meaningful research.
3. Data practitioners should value peer evaluation highly because their professional reputation and future opportunities depend on maintaining the respect and trust of their colleagues.
4. The unanimous professional criticism in this case indicated a clear violation of established ethical standards that any aspiring researcher should take extremely seriously.
5. Professional peers' evaluations are especially meaningful because they come from those who face similar ethical challenges and understand the practical constraints of the field.

Question 3.10: 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?

The software developer's role in this case raises important questions about technical professionals' ethical responsibilities:

1. The developer had an ethical obligation to consider the intended use of his software and its potential for misuse or harm, rather than simply fulfilling the technical requirements.
2. By creating specialized scraping software for this purpose, the developer became complicit in the ethical violation, sharing responsibility for the subsequent privacy breaches.
3. Technical professionals have an obligation to understand and consider the ethical implications of their work, not just its technical feasibility.
4. The developer should have questioned the legitimacy of the request and potentially refused to create tools that would be used to violate platform terms of service and user privacy.
5. While not a study author, the developer's crucial role in enabling the privacy violation demonstrates how technical professionals can be ethically culpable even in supporting roles.

Question 3.11: How do you think the OkCupid study likely impacted the reputations and professional prospects of the researchers, and of the designer of the scraping software?

The actions taken in this case likely had significant professional consequences for all involved parties:

1. The researchers' reputations in the academic community were severely damaged by their disregard for basic research ethics, likely making it difficult for them to publish future work or secure academic positions.

2. The public nature of the controversy and Aarhus University's explicit distancing from the research likely created lasting negative associations with the researchers' names in both academic and professional contexts.
3. The software developer's involvement in creating tools for unethical data collection could make future employers wary of trusting them with sensitive technical projects or responsibilities.
4. The researchers' dismissive response to criticism and lack of remorse likely compounded the damage to their professional reputations, making it harder to rebuild trust within the academic community.
5. The incident's high profile in data science circles means that potential employers, collaborators, or institutions would likely discover this ethical breach during background research, creating long-term career obstacles for all involved parties.