

A2. Proposal for Training on Research Ethics of Big Data

Jing Qian (jq2282)

As the rapidly growing datasets of human information and fast development of big data analytical technology, it is more and more important and urgent to prepare researchers with big data ethics. Ethical issues can be considered in terms of accuracy, humane treatment, informed participants and so on, but big data make such issues more complicated. In this proposal, an entry level training course on the research ethics of big data is designed.

This course is consisted of five parts, which covers from the concepts of traditional research ethics to particular ethical problems in big data research (like updated consent definition), to practical guidance of conducting ethical research with big data.

In the first part, we review the definition and development of Big Data. Big data has great impact on people's everyday life and it is quite different from traditional data. So it is necessary for researchers to be trained on big data ethics.

In the second part, we provide an introduction of research ethics. Here National Research Act of 1974 and Belmont Report of 1979 are discussed. We also introduce the Institutional Review Board (IRB), which conduct formal process for evaluating the ethics of a proposed research.

In the third part, we present social media which is a widely used big data resource and corresponding ethical issues. Start with the famous Facebook study on emotional contagion (2014), we elicit the ethical problem in big data research with social media information. We introduce the content that people's sharing on social media. We discuss people's understanding of consent and privacy of online information and on the other hand, data policy from social media companies.

In the fourth part, we introduce several rules of research ethics in big data that are widely shared by the scientific community. Following questions are the ones that researchers should check: who own the data? Are the data transaction transparent? Do we get subjects' consent and do they understand what they consent to? Is the privacy of subjects violated? Could the data be open to the public? Then we introduce the ten simple rules for responsible big data research proposed by Zook et al (2017). We also show the ethical codes of ACM and IEEE.

In the final part, we introduce the research ethics in different areas of big data, which helps researchers behave ethically in multiple stages of their study. Here we focus on three aspects, data collection and handling, data algorithms and data practice. In fact, most issues involve with all three aspects. In the data collection and handling aspect, we introduce the case of deCODE genetics, a biopharmaceutical company that abused

subjects' consent to do genes research and propose to set up an Icelandic Health Sector Database (HSD) containing the medical records of all Icelanders. In the data algorithms aspect, we explain that due to bias of sample and imperfect design, algorithms can produce unethical decisions, here we use the Microsoft's Tay chatbot as an example. The data practice aspect includes responsible innovation, programming, hacking, professional codes etc.

SYLLABUS

1. Introduction of Big Data.

- 1.1. Definition of big data.
- 1.2. History of big data.
- 1.3. Impact of big data.

* Tests questions.

- 1) What are the main differences between big data and traditional data analysis?

2. Ethical Foundations

- 2.1. History of Ethical Issues.
- 2.2. Institutional Review Board (IRB).
- 2.3. Ethical Concerns.

* Test questions. (Like questions we have in TC0087: HSP(CITI))

3. Social Media and Ethical Issues

- 3.1. Case study: FB study on emotional contagion (2014).
- 3.2. What and why share on social media?
- 3.3. Data policy and consent.
- 3.4. Problem with privacy in big data research.

4. Rules of Research Ethics in Big Data

- 4.1. Principles of data ethics.
- 4.2. Ten simple rules for responsible big data research. Zook et al (2017).
- 4.3. Ethical codes of ACM and IEEE.

* Test questions:

- 1) Which of the following is not the principles of data ethics?
- 2) Which of the following is ethical questionable research?

5. Research Ethics in Different Areas of Big Data

- 5.1. Ethical issue in data collection and handling.
- 5.2. Ethical issue in data algorithms.
- 5.3. Ethical issue in data practice.

* Test questions:

- 1) The FB study on emotional contagion (2014) violate research ethics in which area?
- 2) The deCODE genetic study violate research ethics in which area?