Opinion

EMBO reports

The Cambridge Analytica affair and Internet-mediated research

Christophe Olivier Schneble[1], Bernice Simone Elger[1,2] & David Shaw[1,3]

In recent years, the Internet has become an essential source of data for research. A vast array of information can be collected via platforms, such as Amazon Mechanical Turk [1] and Survey Tools for specific research questions, or from harvesting social networks such as Twitter or Facebook [2]. Questions about data protection, consent and confidentiality will therefore become increasingly important [3], not only for users, but also for researchers and providers of such research and social media services. The European General Data Protection Regulation (GDPR) [4], with its paradigm of security and privacy by default, is a step in the right direction.

The recent scandal surrounding Facebook and Cambridge Analytica [5] shows that these aspects of security and privacy are often not taken into account. Cambridge Analytica, a British consulting firm, was able to collect data from as many as 87 million Facebook users without their consent. The company gained access to 320,000 user profiles and their friends' data through the "thisisyourdigitallife" app developed by psychologist Alexandr Kogan of Cambridge University, UK, when he sold it to the company. Although the 320,000 Facebook users gave their consent for the app to use their data and that of their friends, the latter were not asked for consent and none consented to passing on their data to Cambridge Analytica. Though the information was anonymized and aggregated, the fact that app users were able to consent to the use of their friends' data is very unusual, both in terms of research ethics and social media terms and conditions. When it turned out that Cambridge Analytica had received this data in contravention of its rules, Facebook demanded that the company simply delete the data, without taking any further action to alert the public or warn users. There is still some ambiguity in the media coverage, and even Cambridge University remains unclear about exactly what happened [6]. Nonetheless, the Analytica affair makes it very clear that Internet-mediated research requires much closer ethical oversight.

In traditional human subjects research in psychology and medicine, the ethical evaluation and approval of research projects at the institutional level have become accepted best practice for ethically correct research. However, the relevant guidelines were often designed for medical research, and the issues are substantively different from those in data science. For example, every participant has to give explicit consent for use of his or her personal data in primary research and in much secondary research. Exemptions can be made for anonymized data and when obtaining consent is disproportionally difficult [7]. But today's data science deals with vast amounts of data from various sources, some anonymized, some de-identified and some fully identifiable. Obtaining traditional informed consent from all participants—which normally requires a personal encounter between research personnel and the participant to inform them about use of their data (whether primary or secondary)—is not feasible because of the extremely high number of participants.

We recently conducted a review of ethical guidelines for Internet-mediated research at the top 10 Universities in the USA, the UK and Switzerland [8]. The results clearly show that only a small minority of academic institutions has developed guidelines for data science. This in turn means that most universities are simply not prepared to perform ethical evaluations of research proposals that make use of vast amounts of data collected from social media, secondary apps and the Internet. In general, individual institutions, or at least major research associations, need to start developing appropriate guidelines. We are fully aware that the complex and fast-evolving field of data science does not make this an easy task. One possible way to overcome this problem is to adopt guidelines that serve more as critical reasoning advice rather than making specific suggestions for a single platform or technology, as those of the Association of Internet Researchers [9]. However, this might also turn out to be problematic as ethical considerations in this field of research need a deep understanding of the legal and technical background surrounding it. If the aim remains to develop specific guidelines for Internet-mediated big data research, what are the most important issues that need to be addressed?

Institutional review boards (IRB) need to pay special attention to several issues that may not be adequately covered by existing guidelines. First, commercial or scientific value is often not obvious at the time of data collection [10]; if there is any possibility that future commercial use of data is possible, this should be mentioned in the initial consent. Second, data which are considered public can turn out to have a private character when being aggregated with other datasets. This calls into question the claim made in many guidelines that consent is only needed when data are private: in some situations, combinations of public data might also lead to data being revealed that participants or identifiable groups (especially if they are vulnerable) would want to be kept private. Third, researchers have to pay close attention to the sites from where data are

1 Institute for Biomedical Ethics, University of Basel, Basel, Switzerland. E-mail: christophe.schneble@unibas.ch
2 University Center for Legal Medicine, University of Geneva, Geneva, Switzerland
3 Department of Health, Ethics & Society, CAPHRI Research Institute, Maastricht, The Netherlands
DOI 10.15252/embr.201846579 | EMBO Reports (2018) 19: e46579 | Published online 2 July 2018

collected from and to the properties of the data: is the data protected/publicly available, and what were the "Terms and Conditions" the users agreed when sharing their information? A related point is that data that are anonymized today might be made re-identifiable tomorrow. Furthermore, Alexandr Kogan also worked outside Cambridge University: thus, another important aspect on the Facebook data breach is academics' other jobs. For example, models developed in a research project might be used commercially owing to technology transfer, which could be ethically problematic.

One important practical issue is that IRBs in many countries are not required by law to review such research. However, while IRBs are more used to dealing with health data, the Analytica scandal illustrates vividly that people care deeply about other types of data that are usually subject to national data protection regulations, even if it is not legally regarded as "sensitive data". If data science is to be conducted ethically, IRBs should not wait for the law to catch up, but should review such studies even if legislation does not mandate this. We also believe that social media companies should take the protection of user's data more seriously and deal with this issue more transparently. Currently, issues of privacy and data protection are listed in the terms and conditions but might not be comprehensible to members of the public. At the European level, the GDPR that came into force on 25 May 2018 demands in Article 7 that "the request for consent shall be presented in a manner which is clearly distinguishable from other matters, in an intelligible and easily accessible form, using clear and plain language" and that "any part of such a declaration, which constitutes an infringement of the Regulation shall not be binding". However, it remains questionable whether the GDPR would in practice prevent the common "click and forget" consent systems common to Internet interfaces. This means that IRBs must remain vigilant regarding the information and consent options used in IMR research, particularly when using secondary data and considering waivers of further consent.

A more prominent and understandable way of presenting those issues, as is common practice in traditional clinical research, would prevent further scandals and make the tremendous amount of data that could be used for research more accessible without harming users, but as part of a trusted partnership triangle of social media companies, users and researchers. Facebook has already introduced a new way for users to control their data in a more user-friendly way; research institutions also need to find new ways to effectively guide researchers towards ethical Internet-mediated research.

Acknowledgements

This publication was produced as part of the National Research Program Big Data, Project Regulating Big Data research: A new frontier, financed by the Swiss National Science Foundation. Project Big Data Research Ethics No. 167211.

References

1. Keith MG, Tay L, Harms PD (2017) *Front Psychol* https://doi.org/10.3389/fpsyg.2017.01359
2. Gosling SD, Mason W (2015) *Annu Rev Psychol* 66: 877 – 902
3. Rothstein MA (2015) *J Law Med Ethics* 43: 425 – 429
4. General Data Protection Regulation (GDPR) (2016) (available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN)
5. Lewis P, Grierson J, Weaver M (2018) The Guardian, 24 March (available at https://www.theguardian.com/education/2018/mar/24/cambridge-analytica-academics-work-upset-university-colleagues)
6. Cambridge University (2018) Statement from the University of Cambridge about Dr Aleksandr Kogan, 23 March 2018 (available at https://www.cam.ac.uk/notices/news/statement-from-the-university-of-cambridge-about-dr-aleksandr-kogan)
7. US Department of Health & Human Services (2009) Protection of Human Subjects, 45 CFR 46 (available at https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html)
8. Schneble CO, Shaw D, Zimmermann F *et al* *Swiss J Psychol*. Under Review
9. Markham A, Buchanan E (2012) Ethical decision-making and internet research (available at https://aoir.org/reports/ethics2.pdf)
10. Cate FH, Mayer-Schönberger V (2013) *Int Data Privacy Law* 3: 67 – 73

