Navigating ethics in the digital age: an interdisciplinary approach

Saba Karim Khan, Instructor, Department of Social Research and Public Policy, Division of Social Sciences, New York University, Abu Dhabi, saba.karim@nyu.edu

In this interview with Dr Michael Zimmer, a privacy and internet ethics scholar, we discuss the challenges of navigating ethics during digital research. In light of global data leak scandals, this conversation explores the ambiguity surrounding informed consent, protecting the privacy of online users and informational risk. We ground the discussion in the context of Dr Zimmer’s projects, The Zuckerberg Files and Mapping Privacy and Surveillance Dynamics in Emerging Mobile Ecosystems in the US and the Netherlands, obtaining insights into how ethical challenges practically unfold during research.

Michael Zimmer holds a PhD from New York University and is Associate Professor of Computer Science at Marquette University. He has worked as a Microsoft Resident Fellow at the Information Society Project at Yale Law School.

(Saba Karim, SK): Your project, The Zuckerberg Files\(^1\), is a digital archive of public utterances of Facebook’s founder/CEO, Mark Zuckerberg, spanning 2004-2020. Over 1,000 full-text transcripts and 200 video files are available for researchers to download, analyse, and scrutinise. Why was a centralized database necessary and how does it contribute to the dialogue about privacy and Facebook?

(Michael Zimmer, MZ): The idea for The Zuckerberg Files started back in 2010 when I was chatting with colleagues at the Association of Internet Researchers\(^2\) conference in Sweden, about how it would be great to study how Facebook’s CEO Mark Zuckerberg talks about privacy, and whether that has changed as the platform has evolved. I started gathering copies of his speeches and interviews, and it dawned on me that this could be a useful archive for other scholars to study. So, in 2013 I launched the digital archive of everything Zuckerberg has said in public, and I’ve been working hard keeping it updated ever since.

Why is The Zuckerberg Files important? The dominance of social networking sites, such as Facebook, sparks unique issues of information privacy, the ethics of sharing online, and the corporate responsibility of overseeing a platform with over a billion users that impacts our daily lives, our elections, and just about everything. By understanding how Facebook’s founder and CEO conceives of his own company’s role in the policy and ethical debates surrounding life online, we will be better suited to critically engage in a dialogue on digital privacy, inform design and policy recommendations, and increase user awareness and literacy.

It has been especially fun to see the archive being used in scholarship, news reporting, and even by artists. I do think we’re able to learn more about Zuckerberg and Facebook’s role in our society based on this ability to analyse and scrutinise his public remarks.

(SK): Speaking of Facebook and privacy, in 2014 the company received backlash for a mood experiment involving tinkering with the news feeds of 689,003 users\(^3\). This suggests that whilst digital age ecosystems offer new research opportunities, ethical concerns have taken center-stage. Do you see a human dimension to the social network project for companies and researchers to consider? Are these at odds with one another or reconcilable?

(MZ): Many in the field of computer science have long recognised the need to consider the impact of the tools and technologies they build on humans, including issues as fundamental as human-computer interaction, but also social and ethical impacts. But it does seem that in recent years things have been changing. There’s been the explosion in availability of large and detailed datasets, plus incredible breakthroughs in storage and processing power, all at increasingly lower costs. This puts the power of “big data” and related data analytics, algorithmic decision-making, and machine learning, at the forefront of our current information revolution, and we’re moving so quickly there’s little time to stop and think about the human dimension. “Move fast and break things” is the new mantra, and companies that slow down to worry about ethics might get passed by – this is a big problem.

---

1. https://www.zuckerbergfiles.org/
2. https://aoir.org/
3. Hill (2014)
The past decade has been riddled with controversies – data breaches, collecting and sharing user data without consent, biased algorithms – and the tech sector needs to rethink how they manage big data in a fair and ethical manner. Some companies are making positive moves, including hiring more social scientists to join their data teams, even hiring “ethics officers” to guide product development. But the challenge will be to ensure these roles are empowered within organisations.

From a research perspective, the growth of research projects relying on big datasets about people’s lives and activities, that can be collected without subjects knowing, is testing ethical frameworks and assumptions traditionally used by researchers and ethical review boards to ensure adequate protection of human subjects. Researchers can access petabytes of transactional data, clickstreams and cookie logs, as well as pervasive data from social networks, mobile phones, and internet of things devices. As a result, researchers studying big data find themselves immersed in a domain where information flows freely but is also potentially bound by contextual norms and expectations, where platforms may oscillate between open and closed information flows, and where data may be user-generated, filtered, algorithmically-processed, or proprietary.

All this brings up a set of incredibly thorny ethical issues related to consent, privacy, fairness, and justice. But while many disciplines have a long history of grappling with research ethics, there’s a new generation of researchers in data science who lack this disciplinary history. Hence, they might fail to fully recognise the need for robust ethical deliberations when all they are processing is “data.” Many of us in the data ethics space have been working hard to try to address these issues.

(SK): Are there instances when relying on hacked/leaked data is not considered unethical, especially when it comes to the potential researchable opportunities it offers?

(MZ): Like most ethical dilemmas dealing with data, this is a tricky one and there’s no clear answer. For example, a deontological approach would suggest a duty to not inflict any further harm or indignity against persons already impacted by hacked or other forms of ill-gotten data. However, a utilitarian perspective might seek to weigh any possible harm against the potential benefits that could result from a researcher or journalist making use of hacked data that might otherwise be unavailable. But I think this is where “contextual integrity” (Nissenbaum, 2004) can help, where the ultimate assessment of the appropriateness of a new data flow is the consideration of how the new practice might impinge on the broader values, goals, and ends of the context itself. Here, if social scientists were using hacked data from the Ashley Madison website to simply better understand what kind of people use dating sites targeted to married people, perhaps that’s not completely at odds with the reason people signed up for an account in the first place. But if that hacked data was being used instead, for example, to predict infidelity among the general population for the purpose of improving targeted advertising profiles, that might fail to align with why people joined the platform in the first place and therefore represents a possible ethical violation.

(SK): In terms of ethical boundaries, the power for researchers to observe and experiment on people online, without consent or knowledge, is particularly problematic. This was evident in the Facebook scandal we have been discussing. How does one get around issues of informed consent, protecting user privacy and combatting informational risk, stemming from the potential harm that disclosure of data can cause?

(MZ): I don’t think you can “get around” those issues. These are some of the largest challenges we face in the research community around big data.

As I started to mention above, many disciplines have a long history of engaging in human subjects research (such as medicine, anthropology, psychology, communication) and have longstanding ethical codes and policies intended to guide researchers and those charged with ensuring that research on human subjects follows both legal requirements and ethical practices, and often ethical review boards are charged with approving, monitoring, and reviewing research involving humans to ensure the rights and welfare of the research subjects are protected. But in the new “big data” era – where individuals increasingly share personal information on platforms with porous and shifting boundaries, the aggregation of data from disparate sources is increasingly the norm – the ethical frameworks and assumptions traditionally used by researchers and review boards alike are frequently challenged and, in some cases, inadequate.
This new set of big data researchers are confronted with a continuously expanding set of ethical dilemmas: What ethical obligations do researchers have to protect the privacy of subjects engaging in activities in “public” internet spaces? Which national or international ethical standards apply when researching global networks, communities, or information flows? How is confidentiality or anonymity assured? How is and should informed consent be obtained online? How do we define “harm” to someone existing online in digital spaces? What are researchers’ obligations in spaces which are governed by platform providers? Can researchers violate a platform’s “terms of service” in order to collect valuable data?

We don’t yet have easy answers to these questions, but the big data research community needs to engage with them head on. I’ve been working with colleagues within the Association of Internet Researchers (AoIR) for over a decade to try to tackle these ethical issues, and we’ve recently released our third iteration of ethical guidelines to try to help researchers make informed ethical decisions. We’re taking steps in the right direction to help figure all this out.

(SK): Your project, Mapping Privacy and Surveillance Dynamics in Emerging Mobile Ecosystems in the US and the Netherlands explores how privacy is conceptualised and implemented across cultural contexts. Can you tell us how it contributes to the conversation on digital ethics, as well as provides practical implications for researchers and designers employing a privacy-by-design framework?

(MZ): This project evaluated how smartphone users think about and enact privacy strategies when adopting and using mobile applications that collect sensitive personal information like health data and private conversations. Partnering with researchers in the Netherlands, we examined how people develop mental models to make sense of the perceived benefits and risks of using mobile messaging, health and fitness trackers, and intelligent personal assistants. Through this multidisciplinary collaboration, we evaluated how cultural and contextual factors influence individual users’ attitudes toward privacy and surveillance generally, as well as their attitudes toward specific facets of mobile technologies and specific actors involved in data transactions. We also considered a range of other factors that influence decision-making processes, including the social costs of using these technologies, as well as the affordances and (un)anticipated byproducts associated with the pervasive use of mobile technologies.

The overall goal of this project was to identify and develop easy-to-understand ways for people to make smarter and better-informed decisions when engaging with technology that collects personal information. Our research suggests users have a hard time understanding the value of their data to companies, advertisers, and other third parties, especially when considering how that data can be merged with other data sources. Through our research dissemination among technical communities, we argue that developers in this space should embrace a privacy-by-design framework when designing apps, wearable devices, and other tools that collect, analyse, store, and share large amounts of user data to sufficiently protect users from possible privacy and security risks. Our research provides new directions for future researchers and designers to integrate privacy-by-design principles into mobile devices; to encourage users to consider the nature of the data they generate, as well as account for policy-driven decisions, since greater amounts of personal information are being collected and aggregated.

(SK): Are ethical breaches in digital research different from those in other areas of research? Can rulebooks from other fields offer useful ethical references to researchers using big data?

(MZ): There are multiple aspects of big data ethics that bring unique challenges. First, the relative ease of collecting large amounts of potentially sensitive data has introduced new ethical dilemmas. In the past, most large scale data-intensive research projects (like genomics, for example) required a large amount of funding and specialised equipment, and typically only seasoned researchers with grant funding were able to engage in big data analyses. These researchers typically had formal training in research ethics and had a lot at stake if they made a mistake. But now, I have undergraduate students who can write a python script and grab tens of thousands of profile images from a dating site or comments from a controversial Reddit page and so on, all in a matter of hours. This relative ease of getting data can lead to ethical issues, like in the OKCupid case (Zimmer, 2016), because things move quickly and researchers don’t always have the necessary training.

---

2 Privacy-by-design calls for privacy to be taken into account throughout the whole process (IAPP 2020).
A second factor is that many of these researchers working in the domains of computer and data science who increasingly make use of sensitive data lack a disciplinary culture of considering their work within the purview of an Institutional Review Board (IRB) and simply feel they’re “just working with data” without an adequate awareness that there is a “human subject” connected to that data, worthy of ethical reflection. Much of the motivation behind the PERVADE Project\(^6\) is to help bridge this gap, and ensure that computational researchers increasingly engage with ethics, even when it seems there aren’t any human subjects involved.

This all leads to a unique need for increased ethics training among computer and data scientists. Historically, ethics hasn’t been a central part of computing curricula, but over the past few years, data ethics has quickly become a prominent part of public discourse, drawing increased attention in computing pedagogy, particularly in emerging areas such as data science and artificial intelligence. This has led to calls for greater integration of ethical content within the computing curriculum, but challenges remain on how to best make this happen. Should all computing and data science programmes require a stand-alone ethics class, or should ethics be embedded across the curriculum? Can computer scientists “teach” ethics, or should philosophers be brought into the computer science classroom? We are faced with finding answers to these challenges if we want to ensure computational and data-rich research can continue, but in ethical ways.

**Conclusion**

The issue of ethics using big data, whilst gaining traction, still relies heavily on the “context” of the empirical inquiry for setting ethical parameters. Whilst standard ethical guidelines can be developed, the examples provided by Dr Zimmer illustrate how it can be challenging to apply identical sets of ethical standards to every big data study. Given the immediate access that big data offers, researchers must be cognisant of the risks and threats involved to human subjects. Regular and comprehensive ethical training, developed specifically with big data research in mind, is a useful starting point to make progress in this field.

**References**


---