Identifying Key Industries with Ethical Issues Regarding Data Privacy

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Rushil Korpol

Fall 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Introduction:

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy(IBM Cloud Institution, 2020). Massive amounts of data are fed into algorithms to produce models that serve a wide array of purposes, from predicting the weather to suggesting a new item to buy on Amazon. The use cases of machine learning are endless, but the key point is that they require significant amounts of data. Oftentimes, this data is not ethically collected and comes at the expense and ignorance of the average individual.

A lot of ethical concerns arise with regards to the deployment of machine learning models, particularly in the spaces where personal information is heavily used such as in medicine or insurance. Oftentimes, people are not made aware of what data they are giving up to companies who can later sell that information or use it for themselves. The usage of private user data must be thoroughly regulated to preserve security and safety of people's information, and to prevent biases that may emerge from the use of such data. Simo(2001) expands on some of these concerns, stating that, "The ability to accumulate and manipulate data about customers and citizens on an unprecedented scale may give big companies with selfish agendas and intrusive/authoritarian governments powerful means to manipulate segments of the population through targeted marketing efforts, perform social control, and hence possibly negatively impact the course of our democracies, or do all sorts of harm" (pg 17). This is already seen with China's implementation of their social credit system, where citizens are constantly monitored and penalized for engaging in actions disagreeable with that of the state. In this paper, I will present the ways in which companies in various industries utilize data. I will analyze these use cases and identify the positive and negative impacts, as well as highlight the various ethical issues invovled. Finally, I will explain what protections currently exist around the world to prevent misuse of data, and where this might be lacking.

Problem Definition: The Gap in Understanding Data Privacy

Machine learning has slowly begun to trickle into virtually every industry, and with its wide scale adoption comes a greatly increased desire for data. Brynjolffson and McElheran(2017) explain, "New digital technologies have vastly increased the scale and scope of data available to managers. We find that between 2005 and 2010, the share of manufacturing plants that adopted data-driven decision-making nearly tripled to 30 percent"(pg 1). The adoption of such decision making requires vast amounts of data and results in the creation of a new network with data at the center. As a result, it is essential to pay attention to where this data is coming from and how it gets utilized.

Data has become a major commodity in just the past decade alone, and it is growing at a rapid rate. In fact, according to Forbes, data interactions, which are the creation, copying, and consumption of data, increased by 5000% from 2010 to 2020. It is estimated that the entire big data analytics market is worth 49 billion dollars, and believed to grow by 12% every year. The usage of data has saved corporations billions of dollars; for example, Netflix uses their customers' browsing data to provide recommendations, which has allowed the company to save over 1 billion dollars in customer retention. The benefits of using data to analyze consumer behavior is so clear that over 97% of organizations are investing in data analysis and artificial intelligence in some way. Clearly, data is an extremely important and powerful resource, and as a result it needs to be managed carefully. It is believed that poor data quality has cost the United States over $3.1 trillion dollars annually since 2020. Poor data quality is especially dangerous because it can lead to a misrepresentation of peoples' desires, which would cause companies to act against what people actually want. Additionally, unregulated data protections allows organizations to misuse the information that they have to produce systems, such as with the Chinese social credit system, that exploit people. (Petrov, 2022)

In industries such as medicine and insurance, personal information such as name, address, birthdate, and more is often used as data. Although this has allowed for great progress in these fields, such as with machine learning models being able to predict whether patients have certain diseases, it has come at a cost. In fact, according to Vayena(2016), a recent survey conducted in the United Kingdom revealed that 63% of the population was uncomfortable with their personal information being used for machine learning systems. Even more importantly, it is necessary to properly secure consent to use information, as only then could the highest quality of data be achieved. This is essential because if data is not held to a high standard, then you get, as Vayena puts it, "garbage in and garbage out"(pg 2). Bias emerges with poorly representative data, and this could only serve to harm certain demographics. One example of the current lack of regulation regarding data usage was with Cambridge Analytica during the 2020 elections. As Miller(2019) details, "Cambridge Analytica was a firm that used machine learning processes to try to influence elections in the US and elsewhere by targeting 'vulnerable' voters in marginal seats with political advertising. They had access to the personal information of millions of voters, and developed detailed, fine-grained voter profiles that enabled political actors to reach a whole new level of manipulative influence over voters”(pg 1). This is but one incident where the average individual gets exploited through the misuse of their own data.

As it is now, the protection and ethical usage of data is often overlooked. Consumer data is amassed and sold in large scales by companies around the world, with little to no repercussions. The research paper “Ethical Challenges Posed By Big Data” (2020) highlights this mindset: “In general, it is believed that there is less of a need to protect publicly available information. This has resulted in participants being left unaware of the use, or purpose of use, of their information”(pg 1). For example, many websites today require a user to fill out information about themselves, and in some cases provide their social security number. It is oftentimes required for things such as renting a car, scheduling a doctor’s appointment, or securing an apartment. While it can be understood that social security numbers provide an easy way for organizations to keep track of people, their widespread use also becomes a point of concern. As Darrow and Liechtenstein(2008) put it, “How secure would one feel if they gave the key to their home to every government agency, health care provider, credit card company, and other business organization with whom they have some relationship? What would one do if a copy of this key could be located, inexpensively or even for free on the internet? Yet this is exactly the system that has been created via the use of the social security number as a password that can provide the holder with access to an individual's financial resources, private health information, and more. Worse yet, unlike locks which can be changed if a key is lost or falls into the wrong hands, the social security number is virtually unchangeable.” Their paper reveals a clear flaw with modern practices by companies about requesting data, which ultimately leads to issues such as identity theft.

Identity theft is but one issue when companies are asking users to provide them with data. In the mobile app industry, there is a large issue with apps collecting data from their users without consent or any information that this is happening. People often view phones as an extension of themselves; they contain numerous personal information from fingerprints to geo-location and even vocal data. Apps are able to covertly gain access to this information and send it to their own servers. For example, a study conducted by researchers from the University of Calgary, UC Berkley, and the IMDEA Networks Institute in Spain(Joel et. al. 2019), found that the Shutterfly app was sending its users' geolocation data back to its own servers. The study analyzed a total of 88,000 apps, and found many of them were covertly acquiring data from their users. This included even large apps with millions of active users run by companies such as Disney and Samsung.

However, new legislation around data regulation and protection has started to become enacted. The GDPR (General Data Protection Regulation) is a data protection law enacted in the European Union on May 25, 2018 ("What is GDPR", 2022). It covers a wide range of rights and regulations regarding data privacy, processing, and controlling. Some key points include a requirement of user consent to process their data, the right to access and erase data whenever desired, and transparency with where the data is stored and how it is utilized. The GDPR also applies to all companies that process any EU citizen's data, even if those companies are not located within the EU. Unfortunately, the GDPR is far-reaching and broad in nature, making it a challenge to effectively enforce it.

Yet the GDPR has moved one of the titans of the tech industry: Apple. Apple has enacted a new privacy policy with IOS update 14.5 which will now prompt users for permission for apps to track their activity across other apps and websites. This allows users to decide whether they want to consent with apps collecting their data. Apple even goes to mention how, "Some apps have trackers embedded in them that have more data than they need – sharing it with third parties like advertisers and data brokers. They collect thousands of pieces of information about you to create a digital profile that they sell to others."(Wamsley, 2021) This is a large step up from when users had to manually go into their privacy settings to turn off such tracking.

Despite these new policies in place, there is still a lot to be done in order to ensure an ethical use of data. For one, android phones do not implement Apple’s security policy and are still rife with apps secretly acquiring data. Furthermore, the sale of data needs to be sufficiently regulated, and individuals should be able to decide if they want to allow their data to be commodified in such a way. Finally, when organizations utilize data, they need to be aware of what biases exist in the data in order to ensure there are no ethical concerns with the conclusions from analysis.

Using Actor Network Theory to Understand Ethical Data Usage:

The framework that I will be using to analyze the ethical dilemmas with regards to data privacy and machine learning is Actor Network Theory(ANT). ANT is a sociotechnical theory developed by Bruno Latour, Michael Callon, and John Law in the 1970s that shifts away from the dominant thought process at the time: technological determinism. Technological determinism is a theory that believes technology is what drives society to change, and society adapts to revolutionary technologies by undergoing major transitions. ANT, on the other hand, involves analyzing complex systems as webs of interconnected actors, which are any component in these systems. Actors can range from individuals to organizations and even cultural norms. Any influential element in a system can be considered an actor. ANT is novel in that it considers even nonhuman elements as actors. In Latour's work, "Where are the Missing Masses? The sociology of a few Mundane Artifacts", he describes that to properly identify what role technology plays within a network, it is vital to examine the human work that is replaced(Latour, p.154-155). He brings up several examples such as a door on a wall, where the door opening and closing replaces the human work of manually sealing and unsealing a hole in a wall. This idea can also be extended into more complex technologies, such as with the role of data aggregation replacing the human work of manually asking people questions, writing down information, and storing it so that it can later be compiled and analyzed. With regards to data, by understanding what work is replaced through utilising data, the motivations for doing so can be revealed.

Additionally, I will be using Arnold Pacey's Triangle to organize the various actors into 3 separate groups: technical, organizational, and cultural. Sociotechnical forces can influence and be influenced by the culture of the time and place, the manner in which an authoritative institution behaves, and the technical processes that lead to and from these sociotechnical forces (Neeley, p. 40-42). Through this organization, new insights can be gleamed by digging into the interplay between actors of different groups. An example of Arnold Pacey's triangle is depicted below:

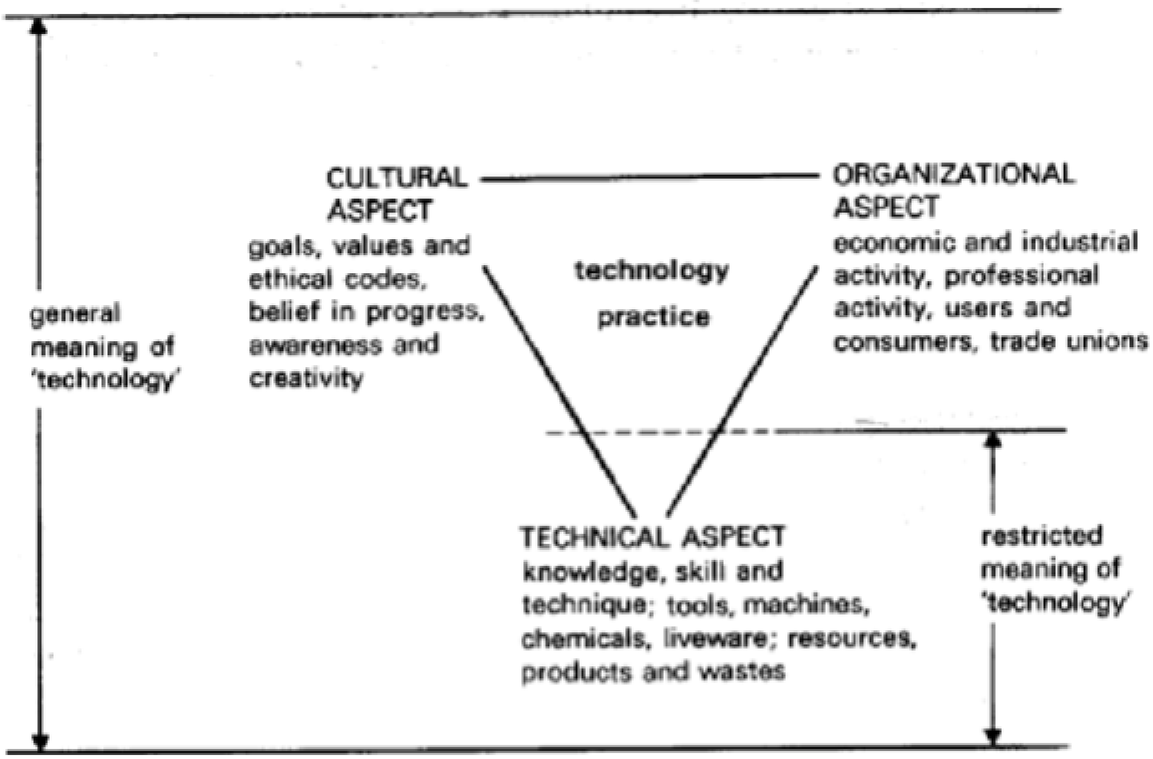
CULTURAL ASPECT
goals, values and ethical codes, belief in progress, awareness and creativity
ORGANIZATIONAL ASPECT
economic and industrial activity, professional activity, users and consumers, trade unions
technology practice
TECHNICAL ASPECT
knowledge, skill and technique; tools, machines, chemicals, liveware; resources, products and wastes
general meaning of 'technology'
restricted meaning of 'technology'

Figure 1: Arnold Pacey's Triangle for Organizing Actors in a Network(Pacey, 1996)

This framework is useful because it will be able to identify specific actors that are involved and their motivations. In "Ethical Issues in Big Data Analytics: A Stakeholder Perspective(2019)", the authors assert that,"We view big data analytics as interactions among stakeholders (individuals, organizations, and society) (Zuboff, 2015). The various interactions between stakeholders may not equitably distribute big data analytics' costs and benefits"(pg 721). I will be using ANT to analyze the relationship between stakeholders across various industries to identify current motivations for utilising peoples' data as well as the costs and benefits associated. In particular, I will be looking at the points of separation between individuals who consist of the general populace, and large monolithic organizations who are harvesting their data, as well as the underlying cultural norms that enable this interaction. I will be analyzing 2 scenarios, one being how Amazon effectively uses data to provide useful recommendations to its customers, and the other how Cambridge Analytica misused social media data to influence voters in the 2016 US election., For each industry, I will dive into what data is currently being utilized, how it is used, what protections exist, and what room for exploitation is possible and the consequences of such. The goal is to provide a scenario in which there are clear benefits to using customer data as long as it is regulated(Amazon), and where there are very large negative impacts(Cambridge Analytica).

Analysis and Results:

Actor networks were created and individual actors categorized based on Pascal's TOC triangle for each of the scenarios. An overview of the two networks will be provided individually first, and then finally comparisons will be drawn.

The actor network for Amazon's use of data in the Commercial industry is shown below:

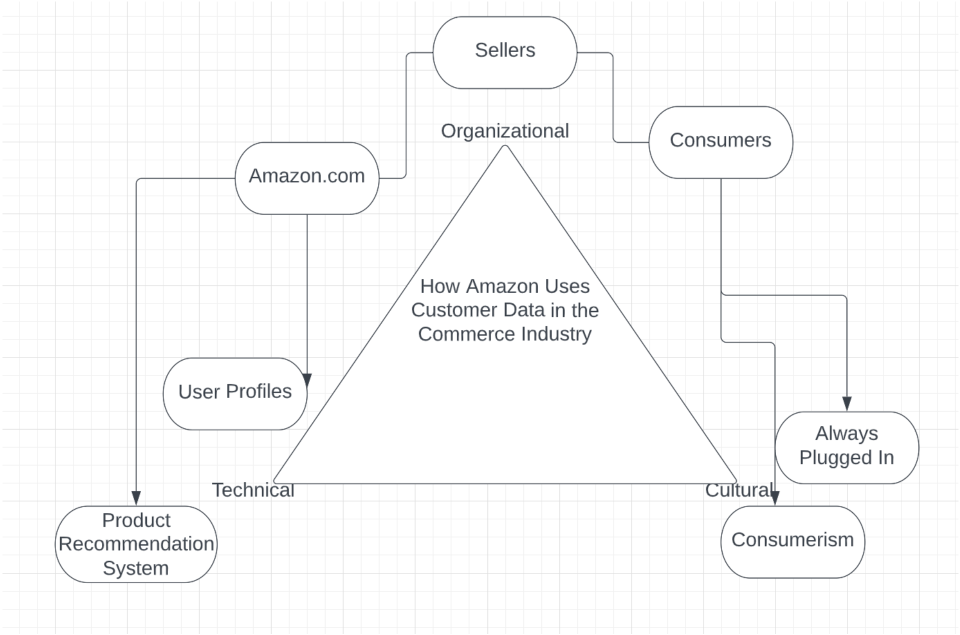

Figure 2: Actor Network for Amazon's Usage of Customer Data(Created by author)

Amazon is a major global company that manages a massive online marketplace found on Amazon.com. Amazon collects extensive information on how users interact with the various elements of the site. They then use this data to recommend new products to consumers as well as provide valuable insight for sellers to allow them to optimize advertising and production strategies(Walter, 2018). These strategies are effective due to a global increase in consumerism as economic prosperity has increased around the world. Additionally, due to access to the internet becoming more ubiquitous, it is easy for many people to quickly scroll to amazon.com and pick out things that they would like to buy.

These recommendations can oftentimes be helpful, as Brent Smith says, "a single book purchase can say a lot about a customer's interests, letting us recommend dozens of highly relevant items"(2017). They allow customers to find new content, but can also remind customers of items they typically include in their shopping cart every week but may have forgot. Clearly, this usage of data is quite beneficial to the customer as it is to the business. However, all this information being collected in one location can prove to be very dangerous. For example, a 2017 study found that Amazon echo user security PINs could be guessed, and cases where "listeners may be able to recover personal details, including payment information" are possible(Haack et al., 2017) .

While there may be cases where using data can prove to provide a net positive, in some cases data can be misused, as seen in the Cambridge Analytica scenario. The actor network is provided below:

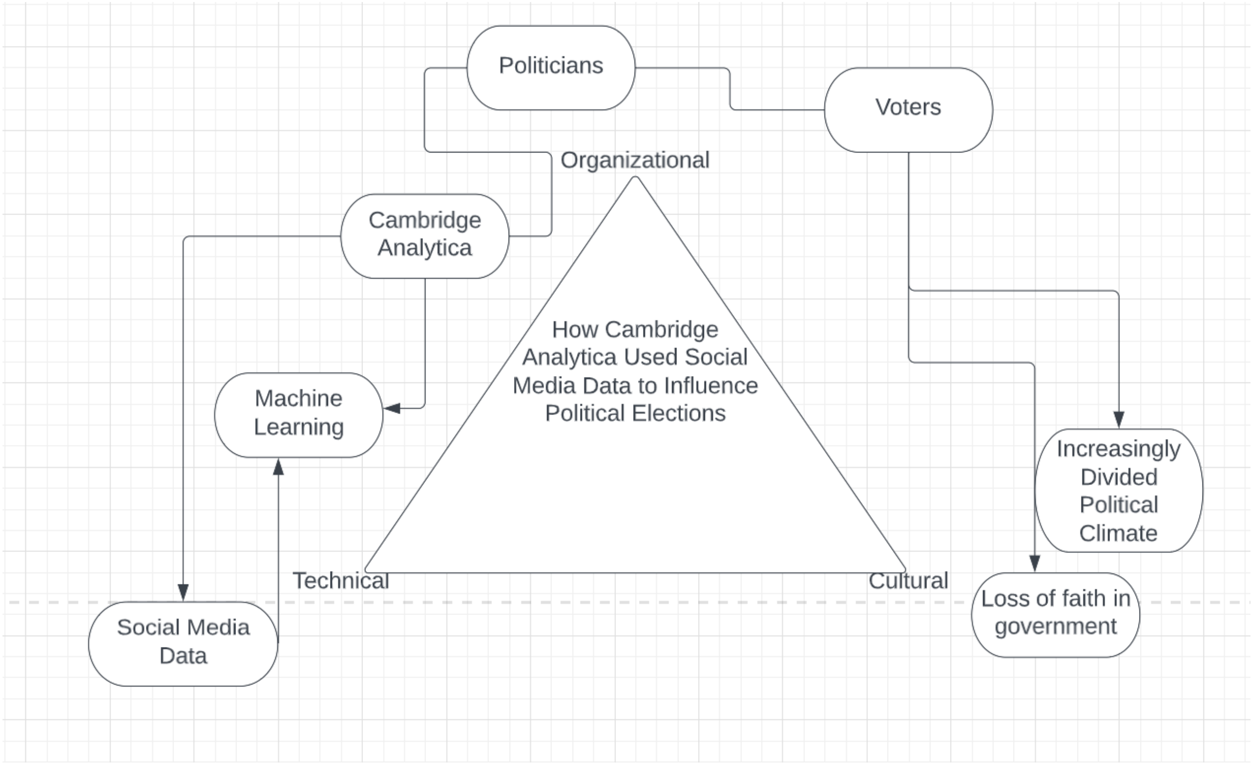
Politicians
Voters
Organizational
Cambridge Analytica
How Cambridge Analytica Used Social Media Data to Influence Political Elections
Machine Learning
Increasingly Divided Political Climate
Technical
Cultural
Loss of faith in government
Social Media Data

Figure 3: Actor Network for Cambridge Analytica Scenario(Created by author)

As mentioned previously, Cambridge Analytica was able to use social media data in order to sway voters to vote for certain politicians by creating targeted posts to slowly polarize opinions. However, it is important to understand that the campaign was effective primarily due to Facebook enabling Cambridge Analytica to gain access to over 87 million users' data without their consent(Boerboom, 2020). This was only perupuated by a growingly divided political climate and a general loss of faith in government culturally that only seeked to increase the gap.

From the two different scenarios, it can be identified that while data can be harnessed to create powerful tools and systems that can be beneficial to individuals, it can also be wielded as a weapon against them. Additionally, it is clear that there is a general complicity with unethical business actions in today's society to a wide host of reasons. The most important reason may simply be due to a lack of education, where people simply do not know that their data is being collected and misused as described in the case of smartphone apps. This lends itself into businesses overextending what they should be doing and results in a great exploitation of the common people. In order to tackle this issue, it is not just important to pass legislation protecting the privacy of people, but also to handle the broader social issues that allowed such behavior to arise in the first place.

Conclusion:

Data is a powerful resource in the 21st century, and as such it should be sufficiently regulated in such a manner that it matches with its scale and possibilities. Although some legislation has begun to arise, such as with the GDPR in the European Union, there is very little currently in the United States protecting and providing oversight for specifically data. Priority should be given to preserving individual privacy and enforcing ethical usage of data. However, as Hands(2018) describes, a balance must be achieved between regulations and progress: "On the one hand, overlooking ethical issues may prompt negative impact and social rejection . . . On the other hand, overemphasizing the protection of individual rights in the wrong contexts may lead to regulations that are too rigid, and this in turn can cripple the chances to harness the social value of data science" (pg 1). An overemphasis of one or the other will only serve to prevent progress and lead to stagnation. There is a clear point in the middle which would allow for companies to ethically use consumer's data, however legislation and cultural shifts would have to begin in order for this to happen.

References:

Berry, E., & Lingard, R. (n.d.2001). Teaching communication and teamwork in engineering and Computer Science. 2001 Annual Conference Proceedings. https://doi.org/10.18260/1-2--9855

Boerboom, Carissa. (2020). Cambridge Analytica: The Scandal on Data Privacy. Augustana Center for the Study of Ethics Essay Contest. from https://digitalcommons.augustana.edu/ethicscontest/18/

Brynjolfsson, E. and K. McElheran (2017) The Rapid Adoption of Data Driven Decision Making, American Economic Review, 106(5), 133-139

By: IBM Cloud Education. (n.d.). What is machine learning? IBM. Retrieved October 14, 2021,from https://www.ibm.com/cloud/learn/machine-learning.

Darrow, Jonathan J. and Lichtenstein, Stephen. (2007). Do You Really Need My Social Security Number? Data Collection Practices in the Digital Age North Carolina Journal of Law and Technology, Vol. 10, No. 1, 2008, from https://ssrn.com/abstract=1699184

Florea, D., & Florea, S. (2020). Big Data and the Ethical Implications of Data Privacy in Higher Education Research. MDPI.

Haack, W., Severance, M., Wallace, M., & Wohlwend, J. (2017). Security Analysis of the Amazon Echo. Retrieved from https://pdfs.semanticscholar.org/35c8/47d63db1dd2c8cf36a3a8c3444cdeee605e4.pdf.

Hand DJ (2018) Aspects of data ethics in a changing world: where are we now? Big Data 6:3,176–190, DOI: 10.1089/big.2018.0083.

Howe Iii, E. G., & Elenberg, F. (2020). Ethical Challenges Posed by Big Data. Innovations in clinical neuroscience, 17(10-12), 24–30.

Joel Reardon, Álvaro Feal, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, and Serge Egelman. (2019). 50ways to leak your data: an exploration of apps' circumvention of the android permissions system. In Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19). USENIX Association, USA, 603–620.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (ed.), Shaping Technology / Building Society: Studies in Sociotechnical Change (pp. 225-258) . The MIT Press .

Miller, S. (2019). Machine Learning, Ethics and Law. Australasian Journal of Information Systems, 23. https://doi.org/10.3127/ajis.v23i0.1893

Neeley, K. Toward an Integrated View of Technology. In Beyond Thintelligence: Toward an Integrated View of Technology (pp. 37 - 45).

Pacey, A. (1983). Innovative Dialog. In The Culture of Technology (pp. 137–159). Cambridge, MA: MIT Press.

Pancake, C. M. (2020). New ways to think about CS Education. Communications of the ACM, 63(4), 5–5. https://doi.org/10.1145/3382126

Simo, H. (2021). Big Data: Opportunities and Privacy Challenges. arxiv.org. Retrieved 2021, from https://arxiv.org/pdf/1502.00823.pdf.
9

Someh, I., Davern, M., Breidbach, C. F., & Shanks, G. (2019). Ethical Issues in Big Data Analytics: A Stakeholder Perspective. Communications of the Association for Information Systems, 44, pp-pp. https://doi.org/10.17705/1CAIS.04434

Vayena, E., Gasser, U., Wood, A., O'Brien, D. R., & Altman, M. (2016). Elements of a New Ethical Framework for Big Data Research. Washington and Lee Law Review Online, 72(3).