DATA COLLECTION: WHY IS IT SEEN AS A PROBLEM?

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

Maximilian Dawkins

Spring 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

S. Travis Elliott, Department of Engineering and Society

INTRODUCTION

According to Nicole Martin, Americans collectively use about four and a half million gigabytes of data each minute (Martin, 2019, p. 1). To put that into perspective, one byte is equivalent to a letter on a page. With roughly three thousand characters fitting on a single page, the data collected in a single minute would amount to almost one hundred thousand pages. After only three minutes, that is enough paper to reach the moon if each page is stacked on top of one another. In reality, much of the data is more compact than text, but four and a half million gigabytes per minute is still a tremendous amount of data. Furthermore, the internet continues to grow, with usage increasing by almost ten percent of the global population in 2018 (Martin, 2019, p. 1).

The rapid growth of the internet created a tool that can connect over half the global population with each other in seconds. Naturally, such a powerful tool sees heavy use as a vehicle for information for organizations around the world. The four and a half million gigabytes of data being used every minute is extremely valuable to companies in the United States. For example, Facebook looks at its consumers' internet search history and uses that information to tune advertisements to each customer's interests. This is one of many ways data is useful to large organizations. In fact, most artificial intelligence (AI) technologies require large amounts of data to work, and global private investments in AI have gone from less than ten billion to almost one hundred billion dollars from 2015 to 2021 (Stanford Institute for Human-Centered Artificial Intelligence, 2022, p. 152). Also, United States government spending on AI reached nearly one billion dollars in the same year, up fifty percent from 2018. (Deltek GovWin, 2022, p. 6). The need for data has been growing, and new uses for data are continuously being discovered.

There is no doubt that the development of the current data infrastructure has lead to many beneficial technologies. To list a few examples, emails are tracked to help filter out spam emails from inboxes, Google Home and Amazon Alexa devices gather speech audio to improve voice detection, and transaction data is analyzed to detect financial crimes. The benefits of big data are apparent in the daily lives of most United States citizens, but many Americans still have concerns about data collection. In fact, a majority of Americans surveyed by the Pew research center believe that there are greater risks than benefits from data collection (Pew Research Center, 2019, p. 4). This paper will look into why there is such concern around data collection, and it will specifically answer the question "Why is personal data collection seen as a problem?". Actor-Network Theory (ANT), an STS framework developed by Michel Callon, Madeleine Akrich, Bruno Latour, and John Law in the early 1980s, will be used to investigate this question.

ACTOR-NETWORK THEORY

The key concept of ANT is that everything exists within an "actor-network" of complex relationships. The relevant actors to the problem of data collection are companies, the government, consumers and citizens, data engineers and analysts, the data infrastructure, data collection methods, and malicious hackers. Each of these actors have relationships with each other, and they each play an important role in the actor-network. Another important concept of ANT for this analysis is dispersed agency (Hurtado-de-Mendoza et al., 2015, p. 330). In the actor-network, agency is dispersed among the actors due to the relationships between them. Determining how agency is dispersed within the data collection actor-network will be a critical part of answering the underlying question of this paper, "Why is personal data collection seen as a problem?". An important note about ANT to keep in mind is that ANT is more of a methodology than a theory. There are no intrinsic properties of an actor-network other than the

fact that it is entirely contained within the relationships between actors. To this end, the analysis in this paper will follow the methodology of ANT, and examine the relationships of each actor. Through an exploration of the relationships in the actor-network, the answer to the paper's underlying question will become more clear.

HACKERS

The exploration will begin with hackers, one of the more clearly relevant actors to the problems with data collection. The most obvious relationship that hackers have with the other actors in the actor-network is data theft. Data theft has long been a concern for companies, governments, consumers, and citizens. Consumers and citizens do not want their personal data to be stolen because many bad things can happen as a result. For example, in 2013 Target's database was breached, with hackers stealing credit card and debit card information from forty million Target customers (Reuters, 2017). In 2014 data from at least five hundred million Yahoo accounts was stolen in a state-sponsored attack, revealing names, email addresses, phone numbers, dates of birth, and security questions and answers (Volz, 2017). The information stolen in these cases can be used to gain access to consumers' accounts and even credit cards. The consumers also have very little control over their data being exposed in large-scale hacks such as the Target or Yahoo attacks. Once they become a part of the system, the users' data is stored in the organizations' databases, and the users have no control over the security of those databases. In this relationship, the agency lies in the hands of the hackers, the companies, and the government. The hackers are the ones causing the breaches while the companies and government have the resources and ability to increase their data security. The only control that the consumers and citizens have is the decision to give their data to the companies in the first place. In the case

of the government census, citizens cannot even choose not to give out their data because it is illegal to not complete the census.

Aside from data breaches, hackers can influence other actors by generating fake data. The effects of this relationship were seen in the 2016 United States presidential election with the generation of fake news on social media. Fake news stories were used as a way to sway the opinions of citizens in the United States towards voting for one presidential candidate or another. Although social media was not a dominant source of news, it was still an important news source, and it played a significant role in the 2016 election (Crawford, 2017). In this relationship, agency lies primarily with hackers and citizens, since the hackers are generating the data and citizens can choose to believe a particular story or not. In the case of elections, the government is affected as well, but also has little agency since it cannot regulate the generation of fake news. Overall, in the relationships with hackers in the actor-network, hackers are the ones with agency, since they are the ones initiating the breaches of data and generating fake data.

DATA COLLECTION METHODS

The next important actors to consider are the data collection methods. In business analytics, there are several commonly used data collection methods (Cote, 2021). The first methods of data collection are surveys, interviews, and focus groups. With these methods, the agency lies largely with the consumers, since the consumer typically has the choice to take part in them. Also, consumers get to choose exactly what information they present to surveys, interviews, and focus groups. In terms of how data collection is seen as a problem, these methods are not very significant due to the contractual nature of them. The next method of data collection is through forms, which are similar to surveys in that the consumer decides what goes into them, but they are typically mandatory in order to use a service. An example of a form would be

providing personal information when making an account for a website. In this case the agency is partially taken away from the user because they are required to at least present some information. User activity observation, transaction tracking, and company social media monitoring are three more methods of data collection. These methods offer even less agency for the consumer, since they happen automatically in the background while a user is using a website, purchasing items, or interacting with the company's social media. The last method suggested by Cote is online tracking. Of the methods listed so far, this one is the most invasive to the consumer because it no longer only collects data through activities with the company. By setting up cookies on a webpage, companies can track online activity data from outside sites. Cookies can track browser search history, IP addresses, and even details like how fast a user scrolls through a website and where their mouse was positioned on the screen. Cookies leave all control to the companies because many users do not even realize what data is being tracked on them. According to a report compiled by PricewaterhouseCoopers, 85% of surveyed UK citizens do not know of any cookie opt-out solutions, and only 13% claim to fully understand what cookies do (PricewaterhouseCoopers p. 8). Also, 56% believe it is important to know how to delete a cookie, which shows that a majority desire to have control over how their data is collected with cookies.

Another important method of data collection is through smart devices. In 2021, there were over ten billion smart Internet of Things (IoT) devices in the world. IoT devices are devices other than typical computers and smartphones that can connect to the internet and transmit data. Examples would be smart TVs, smart watches, smart cars, or smart speakers such as the Amazon Alexa. Each of these devices collect special data, and a lot of it is collected by the various companies that produce the devices. Smart TVs can track not only the shows you watch, but

some “record and send out everything that crosses the pixels on your screen” (Fowler, 2019a). Geoffrey Fowler, a columnist for the Washington Post, tracked the data collected by a 2017 Chevrolet smart car and said, “on a recent drive, a 2017 Chevrolet collected my precise location. It stored my phone’s ID and the people I called. It judged my acceleration and braking style, beaming back reports to its maker General Motors over an always-on Internet connection.” (Fowler, 2019b). Smart speakers even record and collect audio from users’ homes. Collection through IoT devices is the most important method of data collection for this paper’s discussion of why data collection is seen as a problem, because it is the most invasive of consumers’ privacy. In fact, there are twice as many people who find it unacceptable than find it acceptable for companies to even share data collected from smart home speakers with law enforcement to help with criminal investigations (Pew Research Center, 2019, p. 8).

There are many methods for collecting data, some more invasive than others. People seem more unhappy when more personal data is collected, and when they have less control over the data being collected. Different collection methods offer more or less control to the consumer over the collection process, but they all have one common control factor. The consumer must agree to a privacy policy to use products that collect data. This theoretically should offer full control to consumers, making them feel much better about data collection, but this is not true. Only one out of five adults say they often read privacy policies before agreeing to them, and over a third say they never read them (Pew Research Center, 2019, p. 10). Privacy policies are usually long and tedious to read, and most of them do not allow use of the corresponding product without full agreement to the policy. On top of this, consumers do not even trust companies to adhere to their privacy policies anyway, with a majority saying they are not too confident or not at all confident that companies follow their own privacy policies (Pew Research Center, 2019, p.

10). As reported by Thorin Klosowski, the former chief technologist at the Federal Trade Commission claims that “because this [data collection] ecosystem is primarily hidden from view and not transparent, consumers aren’t able to see and understand the flow of information” (Klosowski, 2021).

DATA AND DATA ENGINEERS

The data itself is also an important actor, especially considering its relationships with data engineers. The main relationship between the data and data engineers is simply that the data engineers use data to make data-based technologies. These technologies usually use AI to make predictions or describe relationships, such as Netflix making TV show recommendations, Facebook predicting which advertisements are relevant for a user, or employers gauging performance of employees. As with any technology, there have been failures with AI models, which have lead to mistrust among consumers. One such failure is related to biased data. Biased data stems from systemic bias within the organization that collects the data. For example, Apple’s credit card, Apple Card, used an AI algorithm to determine credit limits for its users. This algorithm ended up discriminating based on gender, offering much higher credit limits to men than women (Telford, 2019). The reason for this was not that Apple itself or its data engineers were deliberately trying to discriminate, but the data that was used to make the algorithm relied on data that was biased. There has been a history of financial discrimination against women, and a lot of the financial data used in Apple Card’s algorithm will contain this exact bias. The Apple Card mistake, among many others similar mistakes related to data bias, contribute to mistrust towards AI and data collection.

Data bias is an unintended consequence of using past data, but problems also arise from intended functions of some data-based technologies. One of the most influential recent online

technologies is social media, as, according to Matteo Cinelli and his colleagues, “social media radically changed the mechanism by which we access information and form our opinions” (Cinelli, et al., 2021). Social media allows for sharing and rating posts by users about anything they want to say, and on many social media platforms, any user can see any posts from any other user. However, since there are so many posts and humans can only pay attention to so many of them, social media algorithms will suggest content to a user that is similar to content that the user previously viewed (Cinelli, et al., 2021). This leads to an effect known as an “echo chamber”, which is a situation where a person’s beliefs are amplified by repeatedly hearing similar beliefs from those around them. The polarization effects of such echo chambers can be seen in social media’s contribution to political polarization of the two main political parties in the United States. Paul Barrett, Justin Hendrix, and Grant Sims, discuss the influential role that social media has had on United States partisan animosity in a tech review by Brookings Institution (Barrett, et al., 2021).

SOLUTIONS AND CONCLUSIONS

In the United States, data is collected from just about every place that one could imagine, and there are many ways of collecting data, but the main concern for consumers is their lack of agency in the data collection actor-network. Thus, a solution to their concerns should target this lack of agency, but there are many relationships in the actor-network through which consumers lose agency. Outside the control of consumers, hackers expose sensitive data stored by large organizations, which is protected through the resources of the organizations. There is little room there to give more control to the consumers, aside from allowing consumers to store their own data. This solution, however, would be cumbersome, and it would create too many new efficiency problems since companies would need to receive their data from the consumers

individually when it is needed. Another problem is that many consumers do not even know to what extent data is collected on them. A solution to this could be the development of a free tool that tracks what information is sent out from various devices. This too has its issues, though, because it still does not actually prevent the collection of data, and it would require a great deal of work to develop. Many consumers simply do not like having their data stored by companies, but they lack control over what data is collected. To use a product, consumers usually have to agree to a privacy policy allowing for data collection, and for many Americans, it is difficult or not worth the time to even read through privacy policies just to decline them and become unable to use the products for which they apply. Also, once data is collected, United States privacy laws allow companies to sell and share consumer data without notifying the consumer (Klosowski, 2021). These privacy policies are currently the best place to provide control to the consumers, since the companies' ability to collect data depends strictly on the privacy policies. The final proposed solution is to lobby for privacy policies or United States privacy laws that offer more control to the consumers. If successful, this would solve many of the privacy issues that consumers have with data collection. With privacy laws similar to Europe's General Data Protection Regulation (GDPR) law, companies would be required to give consumers the right to access, delete, and control the use of their data. This would leave the control over the storage and use of data to the consumers, but achieving a successful change of this degree to data privacy laws would be difficult. That said, data collection occurs on an enormous scale, so it is unlikely to find a simple solution to the problems it poses to consumers. The proposed solutions represent initial ideas for paths to follow in solving these problems.

There is no doubt that data collection plays a big role in the lives of most citizens of the United States, and there are many complex relationships within the actor-network of data

collection that contribute to the problems perceived by consumers. Analysis with the methods of ANT has brought to light many key insights with respect to the relationships between actors, and the biggest takeaway is that consumers have very little agency in the actor-network of data collection. This is not a surprise, considering the scale of many other actors such as the government, companies, and even the data. However, by looking closely at why agency is dispersed as it is, potential solutions to the lack of consumer agency have presented themselves. As with any technology, it is important for the continued development of data-based technologies that consumers trust them, and with sufficient resources and effort, the proposed solutions provide a path for an eventual improvement in the state of trust towards big data among consumers.

REFERENCES

Barrett, P., Hendrix, J., & Sims G. (2021, September, 27). How tech platforms fuel U.S. political polarization and what government can do about it. *Brookings*. https://www.brookings.edu/blog/techtank/2021/09/27/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/

Cinelli, M., Morales, G., Galeazzi, A., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), 1-8 https://doi.org/10.1073/pnas.2023301118

Cote, C. (2021, December, 2). 7 Data collection methods in business analytics. *Harvard Business School Online*. https://online.hbs.edu/blog/post/data-collection-methods

Crawford, K. (2017, January, 18). Stanford study examines fake news and the 2016 presidential election. *Stanford News*. https://news.stanford.edu/2017/01/18/stanford-study-examines-fake-news-2016-presidential-election/

Deltek GovWin. (2022). *Federal Artificial Intelligence Landscape, 2022*. https://info.deltek.com/Federal-Artificial-Intelligence-Landscape-2022-Summary-GovWin-Deltek

Fowler, G. (2019a, September, 18). You watch TV. Your TV watches back. *The Washington Post*. https://www.washingtonpost.com/technology/2019/09/18/you-watch-tv-your-tv-watches-back/

Fowler, G. (2019b, December, 17). What does your car know about you? We hacked a Chevy to find out. *The Washington Post*. https://www.washingtonpost.com/technology/2019/12/17/what-does-your-car-know-about-you-we-hacked-chevy-find-out/

Hurtado-de-Mendoza, A., Cabling, M. L., & Sheppard, V. B. (2015). Rethinking agency and medical adherence technology: applying Actor Network Theory to the case study of Digital Pills. *Nursing Inquiry*, 22(4), 326-335. https://doi.org/10.1111/nin.12101

Klosowski, T. (2021, September, 6). The state of consumer data privacy laws in the US (and why it matters). *Wirecutter*. https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/

Martin, N. (2019, August 7). How much data is collected every minute of the day. *Forbes*. https://bit.ly/3vL2CGn

Pew Research Center. (2019). *Americans and Privacy: Concerned, Confused, and Feeling Lack of Control Over Their Personal Information*. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/11/Pew-Research-Center_PI_2019.11.15_Privacy_FINAL.pdf

PricewaterhouseCoopers. (2011). *Research into consumer understanding and management of*

internet cookies and the potential impact of the EU Electronic Communications Framework. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/77641/PwC_Internet_Cookies_final.pdf

Reuters (2017, May, 24). Target settles 2013 hacked customer data breach for $18.5 million. *NBC News*. https://www.nbcnews.com/business/business-news/target-settles-2013-hacked-customer-data-breach-18-5-million-n764031

Stanford Institute for Human-Centered Artificial Intelligence. (2022). *Artificial Intelligence Index Report 2022* (5th ed.).

Telford, T. (2019, November, 11). Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *The Washington Post*. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

Volz, D. (2016, September, 22). Yahoo says hackers stole data from 500 million accounts in 2014. *Reuters*. https://www.reuters.com/article/us-yahoo-cyber/yahoo-says-hackers-stole-data-from-500-million-accounts-in-2014-idUSKCN11S16P