A Discussion of the Collection and Use of Personal Data by Corporations

Jamie Bly, Robert Chisholm

Abstract— **The collection of user data by media corporations has become a hot topic in recent years. The rise of targeted advertising and the high profile cases of Cambridge Analytica's political meddling and ByteDance's potential banning have thrust the ethics and legality of this collection and use of data into the public consciousness. As such, it is pertinent to take a deeper look into this issue, and into possible solutions. This paper will delve into the methods media companies use to collect and store user data, such as using cookies and storing profiles of users based on the data they have collected. It will explore the ways in which companies use this data, whether in advertising, content algorithms, or for profit. We will also explore the transfer of user data between companies and firms, and the awareness or lack thereof of the user in this transaction. This paper will explore the harm that can be and has been caused by storing this user data, and the criminal actions that have been done, both by nefarious third parties and by the companies themselves. We will discuss the ethics of big data collection, in regards to the consent of the users of these sites and their knowledge about the collection. The article will discuss how much consent is needed prior to collecting a user's data, and what information these companies should be allowed to collect and store. It will explore the potential solutions to the issues presented, mainly lying in legal rulings on the abilities for companies to use and collect this data. It will observe previous rulings and laws introduced related to this issue to analyse how governments might further tackle this issue. We hope our article will give a comprehensive summary of the ethical issues and risks of the burgeoning issue of data collection and data privacy.**

Keywords*— *Cookies, Big Data, Data Privacy, Data Collection, Privacy, Targeted Advertisements, User Data

I. INTRODUCTION

Since the turn of the 20th century, society has raced into an age of technological innovation, unceasingly pushing the limits of human creativity and ingenuity to further the scope of our technology. From planes to trains, algorithms to automobiles, medicine to machines, society has accelerated into an age of incredible technologies, and has not only maintained that speed, but is continuing to gain momentum. The innovations of technology, and specifically innovations in the digital world, have exponentially increased the speed in which information is shared and accessed across the globe, what would have taken days or months mere decades ago, can now be accessible to anyone with access to the internet in seconds. The digitalization of society has emphasised the importance of data and information in everyday life. Everything we do in the digital world can be tracked, collected, stored and then utilised for hundreds of different purposes.

This collection and subsequent utilisation of data is cause for much debate at a societal level. How this data is stored, where it is stored, how is the data used, who has access to the data, and what data is being collected are all questions of unequivocal importance. During an age where the individual's online presence is greater than ever before, and is continuing to grow, there is more data online than ever before, and companies are taking advantage of this increase in online activity.

Companies that have a focus on data are dominating the Fortune 500, with all of the top 3 companies (Walmart, Amazon and Apple) collecting and utilising user data. Looking outside of the top 3, there are companies such as Alphabet, Tesla, IBM and Microsoft, all of which hold significant data-related operations. Even companies without data-related operations collect user's digital data, from conglomerates like Berkshire Hathaway to pharmaceutical giants like CVS Health.

There are a million different reasons for companies to collect their users' data. Berkshire Hathaway Specialty Insurance collects personal information, as well as "your IP address, referring websites, and unique or returning visitors. This data may be linked with personal information we have collected, such as your name or contact information" [1]. The company will hold this information until it is no longer necessary for legal or regulatory reason, and then will be disposed of or de-identified. Tech companies such as Microsoft or Meta will collect your data to "provide you with rich, interactive experiences" [2], which essentially means that they will utilise the data they collect from your interactions with their product to tailor your future experiences on their platforms. For example, Facebook would see that you hypothetically follow pages about classic cars and F1 racing, and using that information would tailor your feed to reflect that interest. This would mean showing you more car-related posts, and that Facebook would show more ads about classic cars, etc, instead of showing a Sephora advertisement for instance. Other platforms like Instagram, Twitter, LinkedIn, and YouTube have very similar functions, with YouTube stating that their collected data is being used to "improve your experience, like reminding you what you've watched, and giving you more relevant recommendations and search results" [3].

While many companies have started being forthright with their collection and utilisation of their users' data, there are still major issues and concerns surrounding the topic. Incidents like the Facebook-Cambridge Analytica Scandal, the October 2022 Microsoft Data Breach and the 2019 T-Mobile Data Breach are a few of the most recent, and most notorious, scandals. As a generalisation, most data breaches can be placed into one of two categories: the misuse of collected data and the theft of personal information.

Misuse of collected data focuses on the instances where the company, organisation, or a rogue employee have exploited their accumulated databases in ways that are unscrupulous or dishonest. The Facebook-Cambridge Analytica Scandal is the prime example of this, as Facebook users' data was utilised for political profiling, which was not within their terms of use.

Theft of personal data is more about exterior third parties gaining access to user data. These situations can come about in multiple ways, through system breach or physical theft directly from the database. The stolen information and data can then be utilised by the thief for their own purposes, or sold on to another interested party for their use. The T-Mobile breach from 2019 is an excellent example of an incident where the hacker was looking to profit from their gains by selling the stolen information. This specific cache of stolen information was reportedly being auctioned for approximately a quarter of a million dollars [6].

II. LITERATURE REVIEW

This section will overview some of the literature we have researched in relation to this topic. First, however, we will outline our methodology for selecting the literature.

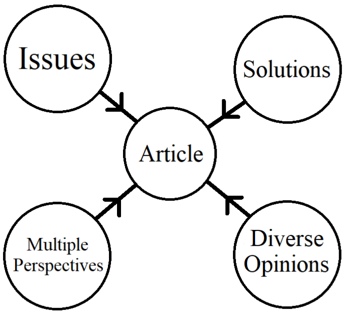


Figure 1. A visualisation of the attributes prioritised when selecting and searching for articles

Our decision process for selecting articles involved attempting to obtain a diverse set of opinions and topics related to this subject. We attempted to find, for example, not just articles on the issues that have been raised surrounding the practice of data collection but also articles on the practical methods and solutions used in the collection and protection of user data. We felt that this would garner us a better understanding of both the issues in this area, and perhaps some potential solutions. While the primary focus of this article is to discuss the inherent problems with user data collection, we hope to effectively summarise all aspects of this field, and have searched for and selected articles accordingly.

A. Protection of Big Data Privacy [4]

This article revolves around the industry it refers to as ‘big data’, defined as ‘‘a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis’’ [4]. In this paper, the processes used to collect and store user data are outlined, as well as the issues inherent in these systems. It then discusses how big data companies attempt to protect the privacy of the users whose data they collect. This reveals some of the concerns that have been raised about big data companies and the collection of user data, and outlines both potential problems and solutions.

B. An Empirical Study of Web Cookies [5]

This article provides an empirical study of web cookies, which is a method that big data companies use to track users who enter and use their sites. The article defines it thusly: “a cookie is a text string that is placed on a client browser when it accesses a given server. The cookie is transmitted back to that server in the header of subsequent requests” [5]. It explains comprehensively why companies use cookies, how they use them to collect data, and what that means for the users who use these sites. This article gives a good insight into one of the most common methods used by big data companies to collect user data, and the potential consequences of their misuse.

C. AI, big data, and the future of consent [7]

This article delves into the growing concern surrounding the collection of personal user data by big data companies, specifically in the realm of transparency and consent. It discusses the colloquial notion of informed consent in other fields (mostly medical), and compares and contrasts them with the ways in which big data companies consider informed consent. This paper discusses the way in which companies can legally “obtain” consent, (such as privacy policies, cookie acceptance popups ect.), and looks into the issues with this form of obtaining consent.

These three articles are able to give a good, broad overview of the many complexities of the collection of personal user data. Some articles merely discuss these issues but we focused on ones that provided potential solutions to these problems in addition to listing them.

III. THE PROBLEM UNDER CONSIDERATION AND POSSIBLE SOLUTIONS

The collection and usage of personal data presents innumerable possible problems and issues, each with their own consequences. The user is entirely at the mercy of the applications and platforms that they utilise, with few possible protections that can adequately protect their personal information and data. This section will be further separated into several segments that highlight individual issues, while also discussing the wider issue at hand. These segments are; Consent and Individual Understanding; Exploitative Use of Data; Negligent Protection of the Data; Excessive Collection of User Data.

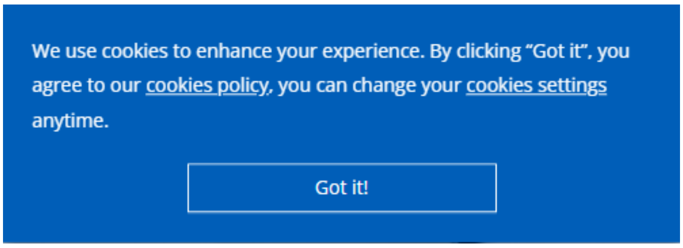
A. Consent and the Individual Users Understanding

The collection, storage and utilisation of data is not innately negative to the user. When used with proper care and respect to the user, it can often greatly improve their experience using the application. As mentioned previously, platforms like YouTube, Facebook, Twitter, and Instagram utilise user interaction with their product to elevate user experience in the future, tailoring their recommendations to their interests and preferences.

While it is a fact that user data can be utilised to better their experience overall, the issue is when the data is used for things that the user has no knowledge of, and thus has not consented to. When entering websites that utilise cookies, oftentimes the user will be prompted to accept that the website utilises cookies, as shown in Figure 2. This prompt is to allow the user to agree to having their data collected, and in return it offers an explanation as to where their information will be utilised.

As touched on previously, the user’s understanding of data privacy is one of the foremost issues in regard to the collection and utilisation of their data. According to a November 2019 study by Pew Research Center, 59% of Americans lack understanding on how private corporations use their data, and 78% confess to having a lack of understanding as to what the government does with their data [10]. Furthermore, approximately 80% of Americans stated that they felt; they lacked control over the data collected, that the potential risks outweigh the benefits of allowing data to be collected, and that they were concerned over how the collected data is utilised [10].

This lack of understanding and trust can be attributed to a multitude of different factors, but there is one factor that if changed would exponentially increase the ease of understanding for the average user. That factor being the fractured nature of the current laws and regulations regarding data privacy and data security in North America.



We use cookies to enhance your experience. By clicking "Got it", you agree to our cookies policy, you can change your cookies settings anytime.

Figure 2. Make-A-Wish Website Cookies Notification

Governments have begun to create laws and regulations regarding the collection and utilisation, but there is no firm all-encompassing law within the United States, and while Canada does have acts like PIPEDA (*Personal Information Protection and Electronic Documents Act*) which encompasses over 75% of provinces and territories, it is not unanimously accepted[8][9]. Having differing laws and regulations between states, provinces and territories creates an excessively large grey area for private corporations. This results in users being subject to different laws depending on where the company who operates the website or application is based. With different laws, regulations, and rules, the average user is left in the dark about how their data is collected, and what it is being used for. Having a prompt that asks for the user's consent before collecting data is all well and good, but for the average user, who doesn't have the time or expertise to understand their privacy policy, giving their consent is verging on irrelevant.

There is an extremely simple solution though. Having one global standard, which all other acts and regulations are built off of, would allow for a lower ceiling to user understanding of how companies use their data. Instead of having 50+ laws and acts that define how companies within North America can collect and use their users' data, there would be one encompassing general law that forms the foundations for regulating data privacy.

Creating a single general standard will make it easier for the average user to understand what is happening, or could happen, to their data. Having differing regulations depending on state or province makes it more difficult for the casual digital citizen to protect themselves. This single change would instantaneously simplify how individuals could protect themselves, which would ease the burden on our most vulnerable digital demographics.

B. Exploitative Use Of Data

In an ideal world, users' data would be exclusively used to the benefit of the user, whether that be through elevating the user experience, or bettering future products. There are hundreds of ways for companies and organisations to utilise their accumulated data in a way that is to the advantage of the user, but there is also the inverse, where there are numerous uses that are not necessarily advantageous for the user. These uses are exploitative in nature, and utilise the data in ways that weren't advertised to the users.

What many companies don't publicise when they prompt the user to accept cookies in order to 'elevate the user experience', is that this includes tailoring advertisements in a way that will encourage users to be more likely to engage with their product and/or service. This could be anything from purchasing more of their products, to renewing subscriptions, to interacting with their services on a more frequent basis. By tailoring advertisements to the user's preference, companies can expand their profit margin, as when faced with advertisements that are more suited to their individual interests, users will be exponentially more likely to engage with the content.

'Targeted advertising', as defined by ADRoll.com, is sending "different messages to different consumers based on what the brand knows about the consumer." [11]. Essentially, this means that if a company or brand can create a user profile of sorts that "demonstrates that it understands what its customers want and need", that the consumer will be more likely to "respond to advertising and engage with the brand." [11]. Research has shown that personalised advertising can boost revenue by 15%, while also increasing the likelihood of repeat purchases [11]. The internet has become the ideal environment and tool for marketers. While targeted advertising is not inherently malicious, and 71% of users prefer targeted advertising [11], the greatest benefactors of the tool are indisputably the companies who utilise it.

Another instance of a company's exploitative utilisation of their accumulated database is the Facebook-Cambridge Analytica scandal. This scandal surrounded the unscrupulous usage of Facebook's accumulated databases, which was obtained through a leak caused by a link between an application developed by Aleksandr Kogan who had access to the Facebook databases [12]. The personal information and data that Cambridge Analytica, a now defunct political consulting firm, had access to was used to create political profiles to influence the 2016 American presidential election [13]. In addition, there are suspicions about Cambridge Analytica's involvement in the circulation of misinformation leading up to Brexit.

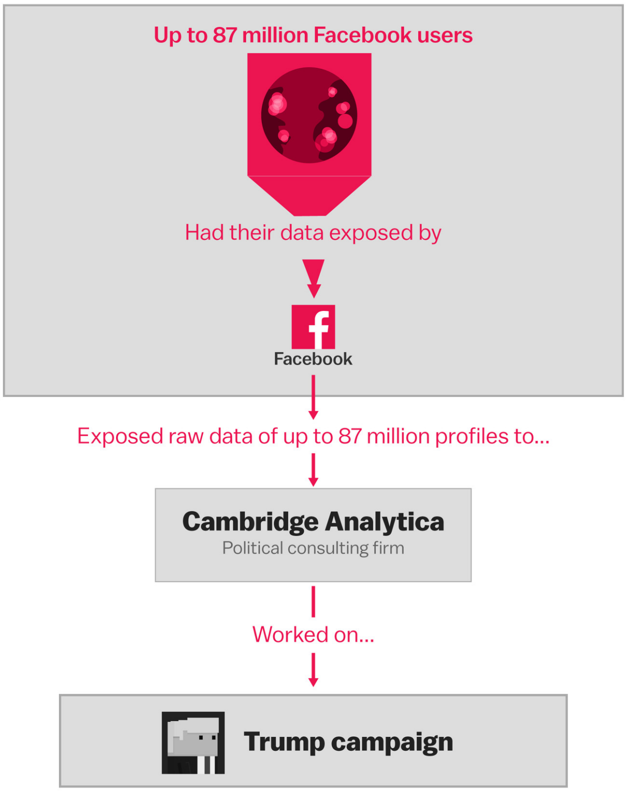

Up to 87 million Facebook users
Had their data exposed by
Facebook
Exposed raw data of up to 87 million profiles to...
Cambridge Analytica
Political consulting firm
Worked on...
Trump campaign

Figure 3. Facebook-Cambridge Analytica Diagram, courtesy of Vox [14]

In the case of Cambridge Analytica, an estimated 87 million users were affected by the leak. These exploitations of the user's data have unimaginable effects on society. Being able to tailor advertising, or campaigning, to the user is an incredibly potent tool, due to the substantial increase in efficacy. Companies utilising users' data must be regulated, especially in a political setting, otherwise society is risking creating 'echo chambers', where individuals are only encountering the same perspectives, opinions and information, and thus develop one dimensional thoughts [17]. Government entities must also regulate targeted advertising to avoid allowing these 'echo chambers' to form, and create damaging environments. Regulating an industry maximum percentage of advertising, (for example 50% of total marketing) that a company can allocate to targeted advertising could be the foundation for reducing the possible detrimental side effects of using user data as a marketing tool.

C. Negligent Protection of User Data

A large issue that comes from big data companies storing a large amount of user data is the potential for breaches and/or theft. Big data companies generally collect data from cookies in two ways. First-party cookies are placed on users' devices by the site's owner, and as such are stored directly in the databases of the site that places the cookies (this does not mean that this data cannot be sold and/or sent elsewhere).

Sites generally use first-party cookies to improve site functionality, such as remembering user preferences or increasing the accuracy of AI recommendation algorithms [5]. Third-party cookies, on the other hand, are placed by other,'third parties', and are generally used to track the user across sites. These are the kind that are generally used by advertising agencies, and are the reason that random websites seem to know what you recently googled or bought off of Amazon [5].

As such, there are several security concerns regarding both types of cookies. Preventative measures must be put in place to stop third-parties from placing cookies in ways that might collect sensitive information [4]. Neglecting this aspect could, and most likely will result in nefarious third parties collecting data from users without their consent.

This is of course ignoring the fact that companies can and will allow nefarious third parties to place invasive cookies without user consent. A famous recent case of this is the aforementioned Cambridge Analytica scandal. In this case, Facebook allowed Cambridge Analytica to collect the personal data of more than 80 million users, all without obtaining consent from the given parties [7]. As such, even when preventative measures are taken, companies can still be willingly negligent with user's data for their own benefit.

There are more issues when it comes to storing large numbers of these 'profiles' created on users in databases. [4] explains: "securing [data] is very challenging. If the big data storage system is compromised, it can be very harmful as individuals' personal information can be disclosed". As such, even if cookies are intended to be exclusively used and accessed by the first-party site owner, this may not remain the case, especially if proper security measures are not taken.

All of these issues can pose a very serious threat to a user's private data security. As such, there is a lot of research into how to protect user data. Firstly, there are applications and tools to prevent third-parties from placing cookies on users of your site. These are generally easy to implement and a good majority of companies use these [5]. If this is not possible however, there are ways to encrypt or muddle the data being collected, so that only the first-party site owner is able to get anything out of the data collected, using encryption keys and various encryption methods [5].

When it comes to storing data, encryption is also incredibly useful, as it can make stolen data much harder to use. Other general data security practices are also obviously useful, such as integrity verification and strict limitations on access and use [5]. It is, however, the nature of cyber-security, that any security system will eventually be compromised, due to new techniques and methods that companies cannot predict. Cases such as the previously mentioned T-Mobile breach are a good example of this [6]. In addition, these security measures do not prevent companies from handing over data to nefarious third-parties willingly. As such, it becomes a question of whether it is necessary for big data companies to collect this much personal information of their users at such a large scale, and if the possible benefits are worth the risk of private data being leaked, stolen or misused.

D. Excessive Collection of User Data

Big data collection can be a useful tool in a myriad of ways. It is the cornerstone of a lot of new innovations in website technology, such as content algorithms and targeted advertisements. However, data collection is often used when it is not necessary. Sometimes it makes certain non-essential processes easier for the site owner, and data is often collected for future use or for things that do not provide a benefit for the customer in any way. Therefore, since this information is incredibly sensitive and extremely difficult to protect from nefarious third parties and/or criminals, should there be restrictions on when and how much data these companies can collect?

This theory of data collection is known as "data minimisation" [16]. It means that when a company decides to collect data, they must ensure that the data being collected is for a specified purpose, and that no more data is being collected than is necessary [16]. Some countries have already attempted to legislate this. The European Union's official stance on this is:

> Personal data should only be processed where it isn't reasonably feasible to carry out the processing in another manner. Where possible, it is preferable to use anonymous data. Where personal data is needed, it should be adequate, relevant, and limited to what is necessary for the purpose ('data minimisation'). It's your company/organisation's responsibility as controller to assess how much data is needed and ensure that irrelevant data isn't collected. [15]

Therefore, companies that are required to adhere to EU regulations must only use data collection in cases where there is no other reasonably feasible option available. Companies cannot collect data for an unspecified future use, and cannot collect more data than is needed for a task. These regulations are a good step forward because, although other countries lack these regulations currently, the EU's stance will hopefully inspire other nations to adopt similar policies. However, this legislation is still lacking in some key areas.

Mainly, the legislation is vague on what is "necessary". While it does specify that if any other method is available to be used for a specific process that it should be used, it does not limit what processes companies are allowed to use. There are many things that require personal data to function, and have no other alternatives, but are not in any way helpful to the consumer. Invasive targeted advertising is expressly done through mass collection of user data, building entire virtual profiles of millions of users, all just to be able to market more relevant items [11].

It could be argued that this kind of process, while unable to be done without user data, is still completely unnecessary and not worth the risk of a data breach. The specific process of creating a "profile" for each user can pose a risk to them even if the data is not breached. Sending advertisements to users about certain products or services might reveal personal details about a user to someone who shares a computer or account.

With current legislation, there is nothing preventing companies from doing these things, and it is incredibly difficult for a given user to avoid having their data used in this way. Further legislation could be passed to prevent particularly private information like this from being stored on company databases, regardless of the feasibility of other options.

IV. CONCLUSION

Ultimately, the field of user data collection is fascinating and multifaceted, but it is also fraught with issues and complications. The data being collected from users is often being taken without their informed consent. Studies show that the majority of people do not know where their data is going and that they are concerned about what it is being used for. As such, it would do well for countries to implement legislation that requires companies to more clearly demonstrate where their user's data is going, and what it is being used for.

Companies have a history of using their users' data for nefarious and exploitative purposes, such as tailoring targeted advertisements to users based off of 'profiles' generated from the data they've collected. Recent, high-profile cases such as the Cambridge Analytica scandal show the way some companies are willing to exploit personal user data.

Proper security measures must be taken when collecting and handling such a large amount of user data. Incompetence in this regard will inevitably lead to data breaches and leaks, allowing unscrupulous third-parties to take advantage of these large pools of user profiles. Government regulations should be put in place to ensure that companies who use 'big data' are putting the proper security measures into place.

Companies also collect much more data than is needed, increasing the risk of harm when a leak or breach inevitably occurs. Companies should stick to the model of 'data minimisation'. Governments could also introduce regulations to enforce this model, like the EU has. They could even go a step further by restricting the ways that companies are allowed to use this data for, such as banning corporations from using personal data to target advertisements.

Overall, while big data collection can be incredibly useful, companies should be more careful and more thoughtful when collecting and utilising user data, and legislation should be put in place in order to ensure that big data will not harm the user.

REFERENCES

[1] "Privacy Policy Statement – United States." *Berkshire Hathaway Specialty Insurance*, https://bhspecialty.com/privacy-policy/#:~:text=We%20will%20only%20collect%20and,accordance%20with%20applicable%20privacy%20laws.

[2] "Microsoft." *– Microsoft Privacy*, https://privacy.microsoft.com/en-ca/privacystatement.

[3] "Understanding the Basics of Privacy on YouTube Apps - Youtube Help." *Google*, Google, https://support.google.com/youtube/answer/10364219?hl=en#zippy=%2Chow-youtube-may-use-data-linked-to-your-account.

[4] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, "Protection of Big Data Privacy," in IEEE Access, vol. 4, pp. 1821-1834, 2016, doi: 10.1109/ACCESS.2016.2558446.

[5] Cahn, Aaron, et al. "An Empirical Study of Web Cookies: Proceedings of the 25th International Conference on World Wide Web." *ACM Digital Library*, 1 Apr. 2016, https://dl.acm.org/doi/10.1145/2872427.2882991.

[6] Cawley, Conor. "T-Mobile Is Investigating a Potentially Massive Security Breach." *Tech.co*, 19 Aug. 2021, https://tech.co/news/t-mobile-massive-security-breach.

[7] Andreotta, A.J., Kirkham, N. & Rizzi, M. AI, big data, and the future of consent. AI & Soc 37, 1715–1728 (2022). https://doi.org/10.1007/s00146-021-01262-5

[8] Office of the Privacy Commissioner of Canada. "Summary of Privacy Laws in Canada." *Office of the Privacy Commissioner of Canada*, 31 Jan. 2018, https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/.

[9] Group, ICGL. "Data Protection Laws and Regulations Report 2022-2023 USA." *International Comparative Legal Guides International Business Reports*, Global Legal Group, 8 July 2022, https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa#:~:text=There%20is%20no%20single%20principal,Code%20%C2%A7%2041%20et%20seq.).

[10] Atske, Sara. "Americans and Privacy: Concerned, Confused and Feeling Lack of Control over Their Personal Information." *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 17 Aug. 2020, https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/.

[11] Lau, Wilson. "What Is Targeted Advertising?" *AdRoll*, 5 Apr. 2023, https://www.adroll.com/blog/what-is-targeted-advertising#:~:text=Targeted%20advertising%20allows%20brands%20to,and%20engage%20with%20the%20brand.

[12] "Facebook: Transparency and Use of Consumer Data." *Facebook: Transparency and Use of Consumer Data Hearing Documentation*, 29 June 2019, https://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Wstate-ZuckerbergM-20180411.pdf.

[13] Confessore, Nicholas. "Cambridge Analytica and Facebook: The Scandal and the Fallout so Far." *The New York Times*, The New York Times, 4 Apr. 2018, https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

[14] Chang, Alvin. "The Facebook and Cambridge Analytica Scandal, Explained with a Simple Diagram." *Vox*, Vox, 23 Mar. 2018, https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

[15] "How Much Data Can Be Collected?" *European Commission*, https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/how-much-data-can-be-collected_en.

[16] "Principle (c): Data Minimisation." *ICO*, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/.

[17] GCFGlobal. (n.d.). *Digital Media Literacy: What is an echo chamber?* GCFGlobal.org. Retrieved April 10, 2023, from https://edu.gcfglobal.org/en/digital-media-literacy/what-is-an-echo-chamber/1/#