An Exploration of Big Tech Data Tracking and its Ethical Considerations

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

Faisal Refai

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

Sean M. Ferguson, Department of Engineering and Society

An Exploration of Big Tech Data Tracking and its Ethical Considerations

Introduction

In 1978, the Electronics Communications Privacy Act was passed. An act which "prohibits the intentional, actual, or attempted interception, use, disclosure, or procurement [of] any other person to intercept or endeavor, to intercept any wire, oral, or electronic communication" (Bureau of Justice Assistance, 2021). While at the time this was not relevant to the media industry, recently the act became prominent in three data tracking court cases: Facebook's ten year class action, Smith v Facebook, and Google's incognito data tracking.

Many social media applications have started to custom tailor their applications to each user's interests. This is very common as machine learning algorithms allow the tracked user data to be put into a model used for features, recommendations, etc. While these features continued to grow, many users were unaware that their data was being collected in order for these features to function. Facebook was one of the first applications to be put in front of the United States government due to data tracking.

In modern technology, more and more companies are starting to implement machine learning algorithms into their products. Many of these algorithms require storage of user data in order to function effectively. As such we need to explore data tracking in social media applications and consider the ethical limitations of where data collection to improve user experience may do more harm than good. Data collection can provide users with more unique and efficient experiences, but the tradeoff is a private entity owning the user's personal data to be sold, analyzed, etc. Without addressing the ethical dilemmas regarding data tracking that have occurred in the past and analyzing them to understand the ethical limits, the use of machine learning and data tracking in technology will continue to be a conflicting area between software

engineers and users. Identifying mistakes made by larger companies in past cases while also focusing on Harvard's defined principles of business ethics, can help engineers analyze the ethical limitations of machine learning and data tracking in technology..

Related Works

TrackCare

The first case study focused on a health care tracking app called TrackCare. The goal of the data tracking was to identify "potential care-seeking events identified through TrackCare, a smartphone application with care-seeking events identified through participant interviews" (NIH 2019). The app would accomplish this through tracking only the final locations of users and not the paths traveled to reach those locations. The study had high levels of transparency asking for consent by all participants and outlining through figures and graphics exactly how the data would be collected. The case was also reviewed by "institutional ethics committees of KEM Hospital Research Centre, Pune and the Centre for Population Health Sciences at the University of Edinburgh" (NIH 2019).

Despite all these attempts to follow ethical standards for data collection, the users still had issues with privacy and confidentiality. The primary concern was despite only final locations of users being tracked, a lot of information can be deduced from location points. This gave the participants a fear of invasion of privacy and a feeling of being "watched". This case is interesting as they followed appropriate procedure and reviews for their study to be considered ethical yet users still felt that their privacy was in danger. While one may argue that this case was ethical as it passed ethics committees and took the appropriate steps. Another may argue that by

collecting data which can be used to determine a user's movement is violating privacy ethics. It shows the difficulty in judging whether a data collection activity is ethical in terms of privacy.

Another interesting outcome of this case study was the language barrier. This study occurred in rural west India and while there were translators to inform individuals about consent and data tracking, some of the more advanced concepts may not have been translated well. Researchers stated "more technical issues such as data encryption and storage and third-party transfer were very difficult to communicate with participants despite good translations of informed consent documents" (NIH 2019). This is something the ethics committees may have overlooked. Yet, one must discuss the ethical implications of a language barrier. Despite consent and transparency being at a high level for this study, it could be invalidated by possibilities of poor translation. One must also use this study to understand that unexpected problems which arise during a study, change the nature of the ethics surrounding the study.

AI based HealthCare App

The AI based HealthCare App differs from the GPS case study in that it is a fictional case study which poses the possible ethical problems which arise from machine learning and AI. The app would be a "a multiplatform application, named Charlie, which utilizes artificial intelligence technologies to make diabetic care easier, more holistic and more accessible. Taking advantage of smartwatches' biosensors to test blood glucose through the skin, the app's algorithms calculate the optimal level and type of insulin for each user" (Princeton 2022).

While there were intentionally many issues regarding consent and transparency placed into the case to create discussion, one interesting aspect of the case was how machine learning was being used. The case stated that users " wanted insight into how Charlie's algorithms worked to construct detailed individual profiles, how it was determined which advice was presented to

individual users, and how the algorithms decided to offer sub-optimal solutions to persons" (Princeton 2022). This provides an opinion of users on data collection regarding machine learning. It shows that users may be open to data collection for machine learning algorithms as long as they are informed of how information is being used and consensually opt in.

Background on Data Tracking in the Tech Industry

Technology companies in recent years have become increasingly put in the spotlight regarding court cases on data privacy. Many users started to realize that the social media applications they were using were tracking information outside of their respective applications to later be sold or stored. While many companies each have their own cases, some of the most notable cases relevant to ethical considerations are Facebook's ten year class action lawsuit for data privacy, Facebook's court case Smith v Facebook regarding medical information collection, and Google's $5 billion lawsuit in the U.S. for tracking 'private' internet use.

Facebook was the first of the large tech companies to be put in front of the US supreme court for data tracking. The court case gained public attention when it was revealed that "Facebook continued to track its users even after they logged out of the social media platform" (NatLawReview, 2022). This was one of the earliest cases which provoked the discussion of the ethics surrounding data tracking done by tech companies.

The case was a class action lawsuit first filed in 2012. The plaintiffs of the case argued that Facebook had not only violated federal and state laws regarding privacy but also the wiretap act of 1978. The official complaint by the plaintiffs was "Facebook unlawfully compiled users' data, including browsing histories, in order to sell their user profiles to third parties for purposes of targeted advertising" (NatLawReview, 2022).

This case was dragged on for several years and was initially dismissed in 2017 until it was reinstated by the ninth circuit court of appeals in 2020. The case was originally dismissed because the plaintiffs struggled to find “alleged concrete and particular harm”. However, after finding proven privacy violations and harm, stemming from sold user data, the appellate court reinstated the case.

While Facebook denied any wrongdoings, the company decided it was in their best interest from a public relations standpoint to settle the case. Facebook is now subject to potentially paying $90 million dollars to the plaintiffs of the case due to the failure to inform their users of data collection in order to profit through targeted advertising. Since the court case started in 2012, Facebook immediately changed their terms and services agreements to account for cases regarding data privacy. While this case was dragged out for many years and was ultimately settled, the ethics of the case probe some questions about user data collection.

The case seemed to set profit as the basis for the line of ethical vs unethical user data collection. Additionally, the case started prior to many machine learning and AI advancements. Machine learning needs to store some form of user data in order to predict actions of a specific user which could be a violation of privacy or the wiretap act. All these topics will prove to be critical as many data privacy cases followed similar paths to this initial case.

Facebook’s Smith v Facebook court case was a smaller court case in the US court of appeals in which “Whether Facebook’s tracking of users’ visits to medical websites violates California and Federal privacy laws” was determined.

The plaintiff of the case argued that the data tracking system with plugins and third party applications that Facebook had in place were in violation of a higher form of data collection, medical information tracking. The plaintiffs of this case alleged that “when they visited certain

healthcare-related websites, Facebook was able to personally identify them and track them through the “share” buttons and “like” buttons embedded on the page” (Epic 2022). Facebook would collect this data in the form of IP addresses of the user, a user’s cookies, and identify the user through a method known as browser fingerprinting. This data was then sold to advertisers for target advertising.

While at this point in time it was no secret that Facebook was collecting information from its users, the more concerning part was it was continuing to do so when visiting medical websites with potentially highly sensitive information.

This case was short lived and did not make it to a higher court of appeals due to the clear wording of Facebook’s privacy policy. The terms of service explicitly states “We collect information when you visit or use third-party websites and apps that use our Services (like when they offer our Like button or Facebook Log In or use our measurement and advertising services)” (Facebook, 2022). The court decided to dismiss the case due to the clear wording of the terms of service which showed that Facebook had consent from it’s users to collect and redistribute such information. However, despite the ruling in favor of Facebook, a discussion can be made regarding the ethics of having such information regarding the policy in a heavy document such as the terms & services rather than a more obvious and frequently seen location.

Google’s court case was similar to Facebook’s 10 year class action lawsuit in that the plaintiff argued that “Google gathers data through Google Analytics, Google Ad Manager and other applications and website plug-ins, including smartphone apps, regardless of whether users click on Google-supported ads” (Stempel, 2020). While the plaintiff agreed that this was expected in normal browsing, they were unaware that this strategy of data collection also occurred in incognito tabs, where the users expected it to be a private session where nothing is

being stored. This led to an in depth analysis of Google's technology and revealed that "Google's software scripts on the website surreptitiously direct the user's browser to send a secret, separate message to Google's servers in California" (CaseText, 2022). There were six elements included in these secret messages; the website being visited, IP address of the user, browser software which is being used, user ID issued by the website to the user, geolocation of the user, and a unique tag which can identify the user. While some of this data is harmless, the geolocation and unique tag for identifying the user were points for concern. The unique tag can connect the data to a specific person and the geolocations track the specific coordinates of a user. Thus the information combined could be used to track a user's real life movements. The plaintiff argued that the collection of this data in the misrepresentation of Google's "Incognito Mode" was a violation of the Wiretap Act. Additionally, collecting geolocations for an application as large as Google which is being used in one form or another essentially means that person could be located and tracked at almost all times.

Google argued to the court that all the claims made by the plaintiff were outlined clearly in their privacy policies and proceeded to explain the reasoning behind data collection. Google started by stating direct quotes from their privacy policies which all users agreed to stating "You can limit the information Chrome stores on your system by using incognito mode or guest mode. In these modes, Chrome won't store certain information, such as: basic browsing information, IP addresses, snapshots of pages visited, cookies, etc." (Google 2022). Google claimed that it gave increased access to its users through incognito mode and that they agreed to the data collection by agreeing to the privacy policies listed in the terms and agreements of using their application.

Ultimately, the court ruled in favor of the plaintiff for two reasons: Google could not show that the plaintiffs consented to data collection and Google could not show that the websites

consented to data collection. Even though Google outlined why and how if collected data in their privacy policy it did not have explicit access to collect user data. This was in violation of the precedent discussed in the Smith v Facebook case in which anyone one who "knowingly accesses and without permission takes, copies, or makes use of any data" was guilty under the CDFA (CaseText, 2022). While the ruling did not apply to the Smith v Facebook case due to the wording of Facebook's privacy policy, it did apply to Google's court case as they did not clearly state the collection of information.

Ethics Discussion of Relevant Court Cases

Defining the Ethical Framework

Each of these cases has ethical implications as they pertain to privacy and human-to-human interaction. However, prior to analyzing the ethics of each case, one must lay out the framework or principles that will be the basis for the analysis. Ethics by nature are subjective and several definitions or key points could be incorporated. For the relevant court cases, this paper will utilize the definition of data ethics set by Harvard Business School.

Harvard Business School defines data ethics as "encompassing the moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals " (Cote 2021). The broad questions that shape this definition are "Is this the right thing to do?" and "Can we do better?". From these questions and definitions, Harvard outlined five data ethics principles for business professionals. Those principles include ownership, transparency, privacy, intentions, and outcomes.

Ownership is the first principle which states that individuals have ownership over their personal data and any collection of such data without consent would be considered unethical.

Transparency outlines the process following consent. Transparency states that once consent is given that individuals have a right to know how their data will be collected, stored, and used. Privacy is another principle which should be practiced throughout the entire process of data collection. Privacy focuses around respecting an individual's personal information. Even if the individual has consented to data collection and the process is transparent, the idea that people do not want personal information to be publicly available should always be kept in mind. Intentions are a principle that should be evaluated at the start of a project but should continually be checked on throughout the process. Intentions are also generally subjective but Harvard defines the principle of ethics as "If your intention is to hurt others, profit from your subjects' weaknesses, or any other malicious goal, it's not ethical to collect their data" (Cote 2021). Outcomes is the trickiest of the principles as even if intentions are good for the project but there is a poor outcome it is known as disparate impact which is unlawful in the United States due to the Civil Rights Act. While you will not know the outcome of your collection until results are complete, there are still ethical ways to continue to prevent harmful outcomes for individuals.

Ethics Discussion of Court Cases

The first of the three court cases was Facebook's 10 year old class action lawsuit which revealed their use of plugins to track their users outside of Facebook. In terms of Harvard's business ethics principles it is clear that Facebook did not follow almost any of the principles. Facebook clearly lacked any transparency and ownership. The plaintiffs lawsuit was on the basis that they had no idea that Facebook was collecting information after logging out. Additionally, Facebook was assuming that the information which was collected was their own to distribute to advertisers. Furthermore, intentions follow similarly to ownership. Facebook's intention was to

collect this data in order to sell to target advertisers to further profit from their users. They did not keep in mind the principle of privacy and the idea that people's personal information could become public by selling to advertisers. Outcome is a principle which could be interpreted in many ways. Since the collection wasn't stopped after the court case but the method of collection did change, the outcome would be increased transparency. Nevertheless, even without the principle of outcome, it is clear that Facebook was not acting ethically during the time of this court case.

Following the class action lawsuit was the Smith v Facebook case regarding storing of medical information by Facebook's plugins. This case is different from the prior court case Facebook was disputing as Facebook had now changed their terms and policy to maintain accountability. By adding in "We collect information when you visit or use third-party websites and apps that use our Services (like when they offer our Like button or Facebook Log In or use our measurement and advertising services)" (Facebook 2022), one could argue that Facebook now accounts for some of the ethical principles outlined by Harvard.

To start, Facebook is now providing transparency to its users about data collection. One could argue that having a small clause inside a huge terms and services agreement isn't the most transparent. However, they are providing the necessary information of their actions and failure to know such a clause is ultimately on the user if they did not read the terms prior to agreeing. The court also threw out any claims of violation of privacy law. While we do not know the exact details of the privacy violations, one would say it's safe to assume that Facebook was not collecting and distributing sensitive information. As far as intentions and ownership are concerned, those did not change between the class action lawsuit and this lawsuit. Facebook still intended to collect the user data to sell to targeted advertisers for profit and while they included a

clause stating collection was occurring they still assumed ownership of the data collected. The outcomes of the case are not determinable as there is limited evidence on whether Facebook's continued data collection had a harmful or beneficial impact on its users. While Facebook still has many things to change regarding data collection from an ethical standpoint, this case shows how Facebook as a technology company is improving to become more ethical from prior cases of ethical & legal issues.

Google's case is unique in comparison to Facebook's cases as it redefines the tech company standards for transparency, intention, and privacy. Unlike the cases for Facebook, Google's court case seems more like a misinterpretation rather than a sneaky attempt to gather information. Unlike Facebook's original class action lawsuit, Google had policies and notification in place to account for data collection. There were clauses in the terms & services and Google even provided a message stating that incognito tabs may still track personal data. Despite all these measures of accountability, Google faced criticism for incorporating data collection in what many users interpreted as a completely private internet browsing session. This critique causes discussions for new standards of transparency. It is now not only enough for companies to incorporate messages and clauses but they must also not make it appear as if users are not being tracked when they are. From an objective standpoint and by definition of transparency, Google acted ethically. However we must account for the fact that appearing to not track data is still not ethical and sometimes following standards by definition may not be enough.

Similar interpretations and changes were seen in the intention principle of ethics. Unlike Facebook, Google did not collect data to be sold to targeted advertisers. Google was collecting data in order to help their own ads program along with the rest of their suite of applications to function better. Google made the argument that its intention was to improve their application,

however this was overlooked by the fact that data tracking was occurring in an incognito tab. Despite correct ethical intentions by business standards, the intentions of the company were overlooked. This probes the discussion about when data collection can and cannot be used. Most data collection occurs in an effort to sell personal information but even with Google altering from such intention they still faced scrutiny. However with what seems like good intentions, the data collection was still problematic.

The final development of ethical principles which occurred from this court case was privacy. Google outlined what information they collected and what it would be used for but that still was not enough to justify the data collection they had done. They provided a lot of transparency regarding the private information that would be collected and what it would be used for. However, users were not concerned about how their data was being used but the fact that their privacy is being infringed just from collection. As such one can start to understand the limits of privacy in terms of what data can be collected and what cannot. The plaintiffs of the case were most concerned with the piece of information which could link the data back to the user. They argued that this was a breach of privacy. While this seems like an easy fix, just collect data without identifiers, it is not that simple. Much of the data collection which occurs is for personalizing recommendations and tailoring content to the user. Without connecting data back to each user, it is impossible to custom tailor such content. One must again analyze the limits of privacy versus machine learning in order to understand the ethical issues surrounding data collection privacy. There are definite tradeoffs between machine learning and privacy as the technology will not work efficiently without the storage of information. The issue lies in that most users want more efficient and personable technology which needs some collection of

unique identifiers. Privacy tradeoffs and personal information collections are important to ponder as applications continue to be made regarding maintaining privacy while utilizing machine

Conclusion

Ethics are a very important yet difficult concept to judge. Due to its subjective nature, it can be easy to justify ethics in favor of one's project or vice versa. However, software engineers and application developers must understand basic ethical principles which should be a baseline for development. Ethical data collection will continue to be an issue throughout the modern technology era with machine learning and AI starting to gain popularity. Yet, one must continue to focus on the ethics of this technology and how it can continue to be improved prior to it becoming mainstream.

References

5 principles of data ethics for business. Business Insights Blog. (2021, March 16). Retrieved March 3, 2022, from https://online.hbs.edu/blog/post/data-ethics

Apte, A., Ingole, V., Lele, P., Marsh, A., Bhattacharjee, T., Hirve, S., Campbell, H., Nair, H., Chan, S., & Juvekar, S. (2019, June). *Ethical considerations in the use of GPS-based movement tracking in health research - lessons from a care-seeking study in rural West India*. Journal of global health. Retrieved March 3, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6596313/

Carolyn Casey, J. D. (2021, July 28). *Google will face class action alleging unauthorized data collection during incognito browsing*. Expert Institute. Retrieved March 3, 2022, from https://www.expertinstitute.com/resources/insights/google-will-face-class-action-alleging-unauthorized-data-collection-during-incognito-browsing/

Data Privacy & Security Alert. (2019, January 7). *Ninth Circuit affirms dismissal of complaint against Facebook for collection of Browsing Data*. Quarles & Brady LLP. Retrieved March 3, 2022, from https://www.quarles.com/publications/ninth-circuit-affirms-dismissal-of-complaint-against-facebook-for-collection-of-browsing-data/

Electronic communications privacy act of 1986 (Ecpa). Bureau of Justice Assistance. (n.d.). Retrieved March 3, 2022, from https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285

Facebook to pay $90 million to settle data privacy lawsuit. The National Law Review. (n.d.). Retrieved March 3, 2022, from https://www.natlawreview.com/article/facebook-to-pay-90-million-to-settle-data-privacy-lawsuit#:~:text=Facebook's%20parent%20company%20Meta%20has,of%20the%20social%20media%20platform.

Lewis, A. (2021, November 3). *The hidden truth behind Google's Incognito Mode*. Reader's Digest. Retrieved March 3, 2022, from https://www.rd.com/article/truth-google-incognito-mode/

LUCY H. KOH, U. S. D. J. (2021, March 12). *Brown v. Google LLC*. Legal research tools from Casetext. Retrieved March 3, 2022, from https://casetext.com/case/brown-v-google-llc

NBCUniversal News Group. (2021, March 22). *U.S. Supreme Court rebuffs Facebook appeal in user tracking lawsuit*. NBCNews.com. Retrieved March 3, 2022, from https://www.nbcnews.com/tech/tech-news/us-supreme-court-rebuffs-facebook-appeal-user-tracking-lawsuit-rcna459

Person, & Stempel, J. (2022, February 15). *Meta's facebook to pay $90 million to settle privacy lawsuit over User Tracking*. Reuters. Retrieved March 3, 2022, from https://www.reuters.com/technology/metas-facebook-pay-90-million-settle-privacy-lawsuit-over-user-tracking-2022-02-15/

Smith v. Facebook. EPIC. (n.d.). Retrieved March 3, 2022, from https://epic.org/documents/smith-v-facebook/

Stempel, J. (2020, June 2). *Google faces $5 billion lawsuit in U.S. for tracking 'private' internet use*. Reuters. Retrieved March 3, 2022, from https://www.reuters.com/article/us-alphabet-google-privacy-lawsuit/google-faces-5-billion-lawsuit-in-u-s-for-tracking-private-internet-use-idUSKBN23933H

The Trustees of Princeton University. (n.d.). *Cases – Princeton dialogues on AI and Ethics*. Princeton University. Retrieved March 3, 2022, from https://aiethics.princeton.edu/cases/