Data and Artificial Intelligence: Mismatch between expectations and uses

By: Diana Garcia[1]

2020

The Author would like to thank the Coastal Virginia Commonwealth Cyber Initiative (COVA CCI) for the opportunity and promoting cybersecurity research

[1] Diana Garcia is a junior at Christopher Newport University, Newport News, Virginia. This paper was supervised by Daniel Shin, COVA CCI Research Scientists, Cybersecurity Researcher, The Center for Legal & Court Technology, William & Mary.

Table of Contents

Introduction

People like to hide behind their phones when it comes to social media. Not every user has their real name or their own photo on display in their social media account. To obfuscate their identities, some users use unusual usernames and profile photos that are divorced from their true identity.

As of December 2, 2020, on Twitter, McDonald's is doing a sweepstake "shave for a free McRib sandwich." (McDonald's) Hundreds of users have posted a selfie of themselves using the hashtag #shave4mcribsweepstakes. Any of the million Twitter users can click on the hashtag and view any uploaded selfies. Selfies are photos that usually contain the person's full-frontal face.

The implication of McDonald's sweepstakes is that users with semi anonymized profiles, such as non-identifiable usernames and profile photos, now have uploaded a picture of themselves and are not anonymous anymore. A photo can contain a GPS location, or the background can reveal the user's address or the user's school logo, which are personal information that you do not want out in the open.

Any facial recognition or profiling technology can utilize this newly available data as well, which can supplement a database full of people's faces. These photos can be used by third parties without the user's consent to improve the facial recognition algorithm and may be used for surveillance purposes. This is a problem to users' privacy and those who can be affected by the technology's bias. There are endless possibilities for what one photo can do.

Artificial Intelligence projects continue to thrive as the increasing availability of computing power makes it more practical to utilize more complex AI models to supplement everyday tasks. (NVIDIA) Artificial Intelligence (AI) enables the machine to make its own

decision with minimal human intervention. Machine Learning (ML) is an AI application that uses large data sets to learn how to perform a given task. Usually, MLs are tasked with labeling a set of data into distinct categories. Deep Learning (DL) is a subset of ML, where it specifically uses artificial neural networks to perform the learning function. Both ML and DL models require large data sets in order to be effective and practical. This need for "big data" can have both legal and ethical implications.

There are federal privacy laws that govern a particular sector, such as the Children's Online Privacy Protection (COPPA), which governs privacy rules intersecting online activities and child end-users, and the Privacy Act of 1974, which governs U.S. federal agencies with respect to personally identifiable information of U.S. citizens and U.S. permanent resident aliens. The Health Insurance Portability and Accountability Act (HIPAA) is widely implemented throughout the healthcare and health insurance industry across the U.S. Unlike in the European Union (EU), there is no comprehensive U.S. federal data-privacy law that focuses on minimizing data collection, securing that data, and informing users about its nature and use. Currently, California's recent privacy law, California Privacy Rights and Enforcement Act of 2020 (CCPA), has filled the regulatory gaps of general data protection for California residents, which is more comprehensive than any privacy laws and regulations of other states or of the federal government. Depending on where you live, each state also has its own state privacy protection laws and regulations.

Technology is advancing quickly but legislation is not keeping up. In other nations, including members of the EU, privacy is seen as a fundamental human right. In the U.S., why isn't data privacy not considered a fundamental right?

When it comes to ethical challenges, data privacy and confidentiality become an issue. Privacy applies to the person and confidentiality refers to the data. Within a world where people are exposed to information technology services that collect volumes of personal information, people are not aware of how their data is being used. The past and potential misuse of user data requires data protection from both the government and private actors. Artificial Intelligence related technologies require volumes of data to function effectively. Gathering and processing personal data can improve the user's experience by making content more personal for the user while allowing the AI engine to perform better and efficiently. Despite the benefits of AI-driven data collection and processing, it should not be hard for people to gain access to their own data and remove it from the system if needed. Ethically, people need to know how and where their data is being used in order for people to maintain their autonomy. Companies using AI technologies have the ethical and legal obligation to not misuse collected personal information from users.

This paper examines one of the issues intersecting the need for big data with the legal and ethical issues of privacy. This paper will address how data is being used beyond the scope of what people anticipated and how little regulations there are for users.

Law and Regulation in the US

There are two major legal definitions of privacy. The first definition stems from the common law perspective, where privacy is the "right not to have one's personal matters disclosed or publicized; the right to be left alone."(Legal Information Institute) The second legal definition focuses on the relationship between the government and its citizens, where privacy is the "right against undue government intrusion into fundamental personal issues and decisions."(Legal Information Institute). According to Dictionary.com, privacy can be defined as "freedom from

damaging publicity, public scrutiny, secret surveillance, or unauthorized disclosure of one's personal data or information, as by a government, corporation, or individual." Overall, privacy can be divided into 4 different classes; information, bodily, territorial, and communications. (Swire and Kennedy-Mayo 21-22) Information includes financial and medical information, government records, and internet activity records. Bodily can include genetic and drug testing, birth control, abortion, and adoption. Territorial examples are work or public space, video surveillance, ID checks, etc. Lastly, communications can be postal mail, telephone conversations, email, etc. Data privacy falls into information privacy. Generally, data privacy focuses on the security surrounding consumer's data and data being used for what data holders are legally allowed to.

The following are some of the major and influential privacy laws. The Children's Online Privacy Protection Act (COPPA) is a federal statute that protects the privacy of children younger than thirteen. The Privacy Act of 1974 allows US citizens and permanent resident aliens to request a change to their records that are not accurate, relevant, or complete from designated federal agencies. California's Consumer Privacy Act, a California state statute, strives to extend consumer privacy protections of California residents from entities with ties to the state. Despite a number of federal and state statutes focusing on privacy protections, there is no comprehensive federal law (so far) that governs data privacy among all sectors of the economy and industries in the United States. In fact, despite growing public concerns regarding cybersecurity, there is no comprehensive federal law that acts as a national data security law covering all major economic and infrastructural areas in the US. Fortunately, the General Data Protection Regulation (GDPR) of the EU contains both data privacy protections and data security regulations across EU member states!

During the 2020 November election, California citizens had the opportunity to consider Proposition 24, which allowed the citizens of California to decide on their own about their privacy. Those against Proposition 24 claimed that this new law undermines and weakens the laws that were already in place and will not enhance Californians' privacy protection. Despite the opposition, Proposition 24 was passed with 56.21% approval, (BallotPedia), opening the road for implementing the Consumer Personal Information Law and Agency Initiative. According to Amends Consumer Privacy Laws Initiative Statute, the summary of privacy proposition is as follows:

- Permits consumers to: (1) prevent businesses from sharing personal information; (2) correct inaccurate personal information; and (3) limit businesses' use of "sensitive personal information"—including precise geolocation; race; ethnicity; religion; genetic data; private communications; sexual orientation; and specified health information.
- Establishes California Privacy Protection Agency to additionally enforce and implement consumer privacy laws and impose fines.
- Changes criteria for which businesses must comply with laws.
- Prohibits businesses' retention of personal information for longer than reasonably necessary.
- Triples maximum penalties for violations concerning consumers under age 16.
- Authorizes civil penalties for theft of consumer login information, as specified

(BallotPedia)

With no federal law on general privacy protection, citizens are only protected by specific federal privacy laws that focus on specific economic/industrial sectors and state laws that may or may not contain general privacy protection. Currently, Virginia does not have an overarching data privacy law that is similar in scope compared to the EU's GDPR or California's CCPA. Nevertheless, pending privacy legislation can help strengthen the privacy protection of citizens in the near future.

Some Virginia house delegates have proposed bills addressing privacy issues. These bills are still pending within the Virginia House Communications, Technology, and Innovation Committee. Some of these bills deal with the care and disposal of customer records, protection in minors online, biometric data, and personal data. Even before the approval of Proposition 24, California was still a leading state with one of the best statutory and regulatory privacy protections for its citizens compared with any other state. In fact, most states only have privacy laws addressing specific sectors of their economy or industry, not an overarching privacy law that provides general data protection of their citizens across all economic and industrial boundaries. States, including Virginia, should study California's approach when state legislatures consider privacy related legislations.

Companies, whose business is mainly handling data of other citizens, will likely face regulatory and legal challenges if more states adopt and enforce more aggressive privacy laws.

Healthcare

HIPAA is a very complex law with lots of moving parts but includes both data privacy and security sections. The privacy rule permits a healthcare provider to use patient data if it is connected to "treatment, payment, and health care operations." However, using the data for marketing purposes or selling protected health information requires explicit authorization. When it comes to people's medical information, it is likely that many people would like to keep that information private. Not everyone would want other people to know what they may or may not have. Patients are more likely to be more truthful to their physicians knowing that their information is kept private. It can also prevent unequal treatment with employers with a veil of medical privacy in place. Trust plays a key role and is an important aspect of the healthcare system. Without trust people have less of an incentive to seek medical care. Over the years,

hospitals and other health personnel have gone digital with their patients' data. With help of the Health Information Technology for Economic and Clinical Health Act, it created an incentive for the healthcare industry to keep their records electronically. HIPAA helps to protect those medical records.

There are already a lot of concerns when it comes to medical ethical issues. Although there is only a little developed legal and ethical research related to AI in healthcare, AI is growing rapidly within the healthcare industry as this emerging technology continues to show its potential. One of the major pitfalls of AI is that biased training data can negatively affect the outcome of the technology. For example, biased data can affect negatively with minority groups, people with disabilities, and people with mental health conditions. Algorithms can also exhibit bias from an inclusive correlation. Despite potential issues with biases, there are methods to remedy these problems. In particular, obtaining data from proper, documented sources can mitigate the effects of biased data sets. Research scientists can also strive for transparency when publishing data analysis to open AI projects for scrutiny, including allowing others to determine whether there is inclusive correlation within the data analysis.

The development of AI in medical technology and the positive changes it cultivates can outweigh the potential negative effects this emerging technology may bring. Anything that improves the medical field can help to save lives. This current flood of AI advancements has been generally characterized by open-programming stages and crowdsourcing. Open innovations empower distant symptomatic frameworks, which speaks to just a little fraction of the advantages that AI offers the clinical world. The engine of machine learning and deep learning frameworks have accelerated the practical applications of AI, but the need for massive data sets introduces potential legal and ethical issues when organizations are collecting medical

information to further develop their AI programs. One of the legal and ethical issues is the problem where collected data is used in such a way that is beyond the intended scope of the data donors.

There are numerous possibilities where data can be utilized in a way that was not intended by the data donors, but here is a hypothetical example. Company X is a medical company that collects donated mammogram scans. Company X states that the donated mammogram scans data improves the AI machines and gets better at detecting cancer within the mammogram scans *and other applications* within the medical field. Donor E is a conservative and religious woman who is glad that her data can help lives and does not ponder nor ask about what other applications can mean. Years passed and Company X gets absorbed by another medical company, Company Y, with services that include cosmetic surgery. Company Y now has access to Company X's donated data and realizes the data can help them develop an AI machine that can help with breast cosmetic surgery. As a conservative and religious woman, Donor E does not approve of this new AI development. Unfortunately, according to HIPAA and the contractual statement she agreed to, her mammogram scans could be used in other applications within the medical field, and thus Company Y is legally not violating anything. Donor E's data is being used in ways she never expected or imagined. With no legal case, Donor E is still distraught. A reasonable ethical perspective may consider Donor E to have been wronged ethically as her donated medical data is being used in ways not only beyond her expectations but also in conflict with her personal beliefs.

Industries

Just by going to a website or signing up for a social media account, there are terms and conditions that users need to click on to agree to be able to gain access. These terms and

conditions end up being pages long. You are not alone when it comes to not taking the time to read everything. When there are updated terms and conditions, once again users are asked to click and agree "yes I have read and agree to the terms and conditions." Do people know what they are agreeing to? The majority of us do not. The following is a brief survey of the terms and conditions of major information technology services.

According to Amazon, AWS (Amazon web services) gives user proprietorship and authority over their content through basic, useful assets that permit them to figure out where their content will be put away, secure their content on the way, and deal with their admittance to AWS administrations and assets for users. Additionally, their systems are intended to forestall unapproved admittance or divulgence of the user's content.

Google assures that it maintains all of the data safes during transit between users' computers or smartphones and its servers. It also states that its cloud infrastructure secures that data and that it does not give any government organizations "direct access" or "backdoor access" to any information. Google indicates that it does not sell consumers' data, but instead uses consumers' information to "make ads relevant" while you're browsing the web. They claim to not consign over any information to advertisers.

Apple Watches continue to gain popularity and dominate the smartwatch industry. With the watch, people can get their heart rate measured with the use of photoplethysmography. Apple's website clarifies that "Blood is red because it reflects red light and absorbs green light. Apple Watch uses green LED lights paired with light-sensitive photodiodes to detect the amount of blood flowing through your wrist at any given moment." Apple has also gained FDA approval for their atrial fibrillation algorithm or Afib and its electrocardiogram, also known as ECG. With the apple watch, it can help detect signs of Atrial fibrillation, a heart disorder. For protection, any

health data gets encrypted and only accessible with the user's permission. Users also have the option to have two-factor authentication turned on. Apple markets itself as privacy-conscious and does not sell user's data. When properly used, people's data can be used to build better and efficient smart systems. Privacy has been one of Apple's biggest selling points. For this reason, Apple has a disadvantage compared to other companies in the area of AI development. For example, according to Apple's privacy page, it states " Your Apple ID isn't connected to Siri, and your requests are associated with a random identifier. Not you." According to sources and research, Apple's Siri is not as efficient as the other services that are available.

Currently, we are in a pandemic. COVID-19 has affected everyone around the world. One important part of preventing COVID is contact-tracing. With Bluetooth, the contact-tracking app is able to communicate with other users who have the app. When a user signifies they are positive, it sends out an alert to anyone who came into close contact with them anonymously. This can create a big impact and open up the country sooner if more people use it. Even though this can create a big impact, there is still a data privacy concern. There is still the concern of how the data can be used in the future, the possibility of it being sold, and breached. This concern is not unfounded, given that the Cambridge Analytica scandal already demonstrated the misappropriate use of collected data.

Cambridge Analytica was previously a political consulting firm and Facebook has millions of users. With the users' data, Facebook approved third-party companies to use the data for academic research. Instead, Cambridge Analytica used the data to help the Trump campaign back in 2016. Christopher Wylie was an employee at Cambridge Analytica and later helped expose the company's efforts. In 2018, Wylie clarified how Cambridge Analytica gathered the data of a huge number of Facebook users, at that point utilizing the information to target

individuals. Users did not merely expose their own personal data but also the data of their own friends as well. Without the friend's user's permission, the user's friend's data could not be accessed by Cambridge Analytica. Users never knew nor expected that doing a survey or downloading an app or anything involved with Cambridge Analytica would lead to their own friends' data being exposed. The CEO of Facebook, Mark Zuckerberg, was later called to testify in front of Congress due to the scandal. Facebook in the end had to pay fines for the incident. Since then Facebook has polished some of its privacy policies and no longer allows friends to share as much information about you. Facebook should have from the beginning been obligated and thoroughly protected the users' data. With no government oversight and no federal laws, Facebook had little motivation to guarantee the safety of their data. If Mr. Wylie never spoke out, no one would have known about this nefarious operation in the first place.

Conclusion

The Internet is a free enterprise, where social media and tech companies have followed an anything-goes philosophy, especially dealing with collected data about its users. This collection and processing of personal data affect millions of Americans. When appropriately utilized, individuals' data can be used to assemble better and effective savvy frameworks. People are ethically obligated to do the right thing, but that is not always the case. Without legislation, private industries are not incentivized to behave righteously, with the Cambridge Analytica scandal being a prime example. Users can unknowingly give away their data, with something as simple as posting a selfie. There are limitless opportunities for what one photograph can do. McDonald's could have conducted the contest in other ways that could have less exposed their McRib participants. With technology and social media growing quicker than legislation laws being passed, it is up to the Industries to create their own policies. At the point when

organizations do look for permission, it is normally through terms of administrative arrangements with excessively long agreements that are brimming with thick lawful language that users are relied upon to "agree" to without comprehension. Information allocation is unquestionably a moral issue. We need new models of information ownership and security that mirror the job data has in the public eye. Even though the current law may allow it, do you think it is ethical for organizations to collect and use data *beyond what was expected* from users?

Bibliography

Albergotti, Reed. “Apple's new watch draws attention to its health-care play.” *The Washington Post*, 16 Sept. 2020, https://www.washingtonpost.com/technology/2020/09/15/apple-event-2020-apple-watch/. Accessed 15 Oct. 2020.

Apple. “Monitor your heart rate with Apple Watch.” *Apple Support*, 18 Sept. 2020, https://support.apple.com/en-us/HT204666. Accessed 15 Oct. 2020.

AWS Amazon. “Compliance.” *Amazon*, https://aws.amazon.com/compliance/data-privacy-faq/. Accessed 4 Nov. 2020.

BallotPedia. “California Proposition 24, Consumer Personal Information Law and Agency Initiative (2020).” *Ballotpedia*, 2020, https://ballotpedia.org/California_Proposition_24,_Consumer_Personal_Information_Law_and_Agency_Initiative_(2020). Accessed 22 Nov. 2020.

Brady, Holly. “Virginia - Data Protection Overview.” *DataGuidance*, 21 Aug. 2020, https://www.dataguidance.com/notes/virginia-data-protection-overview. Accessed 19 Nov. 2020.

Burton, Andrew. “Apple's Privacy Pledge Complicates Its AI Push.” *Wired*, 14 July 2017, https://www.wired.com/story/apple-ai-privacy/. Accessed 23 Oct 2020.

Daniel Schönberger, Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications, *International Journal of Law and Information Technology*, Volume 27, Issue 2, Summer 2019, Pages 171–203, https://doi.org/10.1093/ijlit/eaz004

Dictionary. “*Privacy*”. (n.d.). In *Dictionary.com*. https://www.dictionary.com/browse/privacy. Accessed 21 Oct. 2020.

Google. “Data Privacy Settings & Controls - Google Safety Center.” *Data Privacy Settings & Controls - Google Safety Center*, https://safety.google/privacy/privacy-controls/. Accessed 29 Oct. 2020.

Green, Andy. “Complete Guide to Privacy Laws in the US: Varonis.” *Inside Out Security*, 30 Mar. 2020, https://www.varonis.com/blog/us-privacy-laws/. Accessed 19 Oct. 202.

HHS. “Your Rights Under HIPAA.” *HHS.gov*, 02 Nov. 2020, https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html. Accessed 21 Oct. 2020.

Lee, W. W., Zankl, W., Chang, H. "An Ethical Approach to Data Privacy Protection", Isaca Journal Volume 6, 2016.

Kahn, Stasia & Sheshadri, Vikram. (2008). *Medical Record Privacy and Security in a Digital Environment*. IT Professional. 10. 46 - 52. 10.1109/MITP.2008.34.

Koetsier, John. “Alexa, Siri, Google Assistant: How The Top Smart Assistants Stack Up.” *Forbes*, 08 Aug. 2020, https://www.forbes.com/sites/johnkoetsier/2020/08/08/alexa-siri-google-assistant-how-the-top-smart-assistants-stack-up/#3e2b6c60748b. Accessed 23 Oct. 2020.

Lapowsky, Issie. “How Cambridge Analytica Sparked the Great Privacy Awakening.” *Wired*, 17 Mar. 2019, https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/. Accessed 16 Nov. 2020.

Legal Information Institute. “right to privacy.” *Legal Information Institute*, https://www.law.cornell.edu/wex/right_to_privacy. Accessed Nov. 2020.

Mandal, Subhamoy et al. “Imaging Intelligence: AI Is Transforming Medical Imaging Across the Imaging Spectrum.” *IEEE pulse* vol. 9,5 (2018): 16-24. doi:10.1109/MPUL.2018.2857226

McDonald's. “McDonald's on Twitter: "k giving them away only IF you shave bc beards + McRib sauce don’t mix. post ur clean-shaven selfie on ur public profile with #shave4mcribsweepstakes and @mcdonalds. First 10,000 could get a free McRib via @ubereats. ends 12/2.” *McDonald's on Twitter: "k giving them away only IF you shave bc beards + McRib sauce don’t mix. post ur clean-shaven selfie on ur public profile with #shave4mcribsweepstakes and @mcdonalds. First 10,000 could get a free McRib via @ubereats. ends 12/2.*, Twitter, 30 Nov. 2020, https://twitter.com/McDonalds/status/1333424771552452611. Accessed 4 Dec. 2020.

Newman, Daniel. “Privacy Pros And Cons As Apple And Google Look Into Using Data To Trace COVID-19.” *Forbes*, 22 Apr. 2020, https://www.forbes.com/sites/danielnewman/2020/04/22/privacy-pros-and-cons-as-apple-and-google-look-into-using-data-to-trace-covid-19/?sh=5473d69851fa. Accessed 11 Nov. 2020.

NVIDIA. “AI AND DATA SCIENCE.” *NVIDIA*, https://www.nvidia.com/en-us/ai-data-science/. Accessed Nov. 2020.

Park, Alice. “Here's How Well the Apple Watch Can Detect Heart Problems.” *Time*, 14 Nov. 2019, https://time.com/5727608/apple-watch-heart-study/. Accessed 15 Oct. 2020.

Pesapane, Filippo et al. “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine.” *European radiology experimental* vol. 2,1 35. 24 Oct. 2018, doi:10.1186/s41747-018-0061-6

Robin Feldman, Ehrik Aldana, and Kara Stein, *Artificial Intelligence in the Health care Space: How We Can Trust What We Cannot Know*, 30 *Stan. L.& Pol'y Rev.* 399 (2019). Available at: https://repository.uchastings.edu/faculty_scholarship/1753

Sandra L. J. Johnson (2019) *AI, Machine Learning, and Ethics in Health Care, Journal of Legal Medicine*, 39:4, 427-441, DOI: 10.1080/01947648.2019.1690604

State of California. "General Election: Final Ballot Labels and Titles and Summaries." *CA Elections*, 13 Aug. 2020, https://elections.cdn.sos.ca.gov/ccrov/pdf/2020/august/20172rm.pdf. Accessed 21 Nov. 2020.

Swire, Peter, and DeBrae Kennedy-Mayo. *U. S. Private-Sector Privacy: Law and Practice for Information Privacy Professionals*. Third ed., IAPP, 2018.

Yu, Kun-Hsing et al. "Artificial intelligence in healthcare." *Nature biomedical engineering* vol. 2,10 (2018): 719-731. doi:10.1038/s41551-018-0305-z

Zinolabedini, Darius, and Nikhil Arora. "The Ethical Implications of the 2018 Facebook-Cambridge Analytica Data Scandal." *The University of Texas at Austin University of Texas Libraries*, 2019, https://repositories.lib.utexas.edu/bitstream/handle/2152/80574/AroraZinolabediniThe%20Ethical%20Implications%20of%20the%202018%20Facebook-Cambridge%20Analytica%20Data%20Scandal.pdf?sequence=2&isAllowed=y. Accessed 1 Dec. 2020.