Genetic and Facial Recognition Data and Our Privacy

Jeanne Gerlach Lewis
October 16, 2020

In her exhaustive history of privacy in America, Sarah Igo argues that although privacy has been an American concern since the 19th Century, it has become an acute concern in the present day. The reasons for this are obvious: Emerging technologies and media, novel modes of professional and corporate surveillance and new practices of official documentation have combined to break down previously understood barriers between our private and public selves (Igo: 1-10).

Certainly, the benefits of these advances are self evident. But their costs in anonymity and even autonomy have only recently been widely recognized. Today we are not only struggling to define and defend an acceptable boundary between the public and the private in our contemporary capitalist republic but are in fact challenged to define precisely what we mean by privacy. This endeavor is complicated by the fact that privacy is neither empirical nor logical but rather a symbolic cultural construct. That is, it is a category of meanings that are bipolar and multivocalic. Many of these meanings seemingly contradict each other and are differently emphasized and utilized by different members of our society in different times and places and none of these varied meanings are equally accessible to all our citizens.

The problem of privacy is then a philosophical and cultural problem as well as a social, legal, political and technological one. Any real solution to this problem has to address all these factors. Thus, this is a large, complicated problem and one that is expanding and changing continuously. I will attempt to explore it by looking at the threats posed to privacy by two technologies, genetic technology and facial recognition technology. I cannot explore either in depth in the time allotted but I do hope to give you some sense of the complexity of the more general privacy problem and an understanding of why it is so difficult if not impossibly to solve.

Since Crick and Watson announced that they had determined the double-helix structure of DNA in 1953, the applications of DNA analysis have multiplied rapidly. Among those following directly from the discovery are prenatal screening for genetic diseases, genetically engineered foods, an increased ability to identify human remains, the rational design of treatments for diseases like AIDS and accurate testing to convict or exonerate suspected criminals (History.com Editors: 2/26/2020).

In the past decade genetic technologies have also become a major source for the gathering of personal genealogical and genetic information by individuals. With the American genius for commodifying virtually everything companies like Ancestory.com, GEDmatch, 23 and Me, Map

My Gene and Origin 3N have made it possible for private individuals to submit samples of their DNA to the company of their choice which, then for a price, will map their genome and combine it with information derived from birth records, family trees, newspaper clippings and other public records to create a story about the DNA owners' family.

Sometimes, perhaps most times, the story created through this genetic genealogy technique is satisfying and reassuring to the client. But it can also contain surprises both pleasant and unpleasant. Some people discover a long lost relative or a connection to a famous person. But others may be shocked or offended by what they find out. I have a friend, for example, who discovered that her mother was a bigamist and her father was not her father. In my own family genetic genealogy led to the revelation that my great-grandmother was not the genteel Southern belle of family legend, but rather a lady of uncertain reputation from upstate New York.

The popularity of this practice has made genetic testing companies profitable and allowed them to aggregate massive amounts of data in their data bases. Even a relatively unknown company like Family Tree DNA has a genetic database of about two million Americans while the more popular Ancestry and 23 and Me have enough data between them to identify nearly all three hundred million Americans. (Murphy: 12/23/19; Joh: 6/11/19)

One area where these huge data bases have the potential to be a public good is their use in producing new and improved medical treatments. Since 1996 the HIPPA protocols have prohibited doctors, nurses and pharmacies from sharing health care data with third parties without patients' explicit permission. However, since health care data was digitized in 2009, machine learning algorithms have been allowed to use anonymized health care records and demographic data to improve diagnoses and treatment across all medical fields, yielding earlier diagnoses, better treatments and improved causal understandings.

For example, in May of 2019 Google and Northwestern University teamed up to apply a deep learning algorithm to 42,290 CT scans used to predict the likelihood of lung cancer. The model was able to predict cancer 5% more often with 11% fewer false positives than six experienced radiologists. Human brains are capable of this same kind of deep learning but not at the same speed on a comparable amount of data as a machine. (Frank: 10/2/19)

The data used in this case was owned by Northwestern. But similar usage by other third parties such as research institutions, public health agencies and pharmaceutical companies often look to genetic genealogy companies for data. These companies speak of sharing such data but in fact they sell it, sometimes with clients' permission, sometimes without it. Other pertinent data comes from the well-intentioned practice of individuals uploading genetic information of their own or of loved ones to so-called open-source databases like GEDmatch and My Heritage, seeking support or trying to help others who share the same health-threatening information. The people who do this are often parents who have had their children genotyped and discovered potential health problems, many of which will not be symptomatically apparent until the children age.

Both sources of genetic health data are problematic. Health care data breaches put individuals at risk for identity theft, fraud, the risk of being psychologically profiled, denial of insurance coverage and the risk of elevated insurance premiums. Parents are exposing their nonconsenting children to a lifelong exposure of their data that may have negative consequences at school, work and in the insurance marketplace. (Bela: 1/2/20) Adults who choose to share their genetic data with researchers and pharmaceutical companies risk their own privacy as well as that of their descendants and collateral relatives. Experts agree that any anonymization of data is in fact impossible (Cofone: 4/19/18, Frank: 10/2/19, Colata: 7/23/19) and it is often the case that companies holding this data often do not honor their clients' choices not to share it. (Bamberger, Engleman, Han, Bar On and Reyes: 12/5/19)

The second widespread third-party user of genetic data is law enforcement. This data is used to identify the bodies of unknowns and missing persons as well as to identify and convict criminals, most famously rapist and murderer Joseph De Angelo, the so-called Golden State Killer. DeAngelo was not the first criminal to be identifies through a DNA match but was the first to be identified using genetic genealogy. Law enforcement has long had access to Codis, the FBI's national criminal data base which includes genetic data gathered at crime scenes. It is most useful when investigators have a crime scene sample that they can match to an identified sample in Codis. It is less useful when there is no identified match. What genetic genealogy can do is to bridge the gap between an existing sample and similar but not identical samples in other data bases.

This is exactly what happened in the DeAngelo case. He had left his DNA behind at a number of the rape murders he committed but had never been arrested and given a DNA sample to police. To identify him investigators created a fake profile with DNAmatch, one of the first companies to do mail-in DNA analysis. Looking at its database of clients who had voluntarily submitted their DNA, DNAmatch was able to identify a number of individuals whose DNA varied in ways similar to the sample's and whose genealogies included each other as well as a third cousin, DeAngelo. (Murphy: 7/1/19, Murphy and Levenson: 3/5/20) Shortly thereafter in June 2019, genetic genealogy was used to identify and convict the 1987 murderer of a Canadian couple using the GEDmatch database and a DNA sample from the crime scene that had been uploaded to Codis at the time of the crime.

Genetic genealogy thus became an important new tool for law enforcement. Experts agree that it could revive investigations into 100,000 major violent crimes and reveal the identities of 40,000 bodies, provided of course that commercial databases are willing to share their data. An of course many are, for a price and in some cases with some restrictions. Parabon, a forensic consulting firm charges $3500.00 to build series of interlocking family trees that reveal all descendants from a single pair of great-great grandparents. Law enforcement as well as private individuals can purchase Parabon's services (Murphy: 4/25/19). Family Tree DNA, with a database of two million people, charges the relatives of rape and murder victims $800 to search it, yielding findings that can be shared with law enforcement. The founder of GEDmatch sold his database to Virogen, a commercial forensics company whose major clients are the FBI and police departments.

Other database proprietors are reluctant to share data with law enforcement because to do so explicitly breaks privacy agreements with their clients and because it also exposes them to increasing pressure from law enforcement to lower the bar for disclosure to nonviolent and even minor crimes like vandalism. Even when databases stipulate that they will share data with law enforcement only in cases of rape and murder and only with a court order they have been forced to so by judges sympathetic with law enforcement arguments. Recently Florida judges have granted search warrants for nonviolent crimes and have explicitly allowed searches of all user data whether the clients had chosen to classify it as for use by law enforcement only, no use by other clients or no use of any data (Murphy: 12/23/19, Orenstein: 6/13/19).

Public concern about protection from and access to this type of data by law enforcement has led to Department of Justice Guidelines that limit searches of genetic genealogy databases to violent crimes and the identification of human remains with the following stipulations: Law enforcement must exhaust other investigative means before applying to search a database. Law enforcement cannot "steal" a DNA sample for a database search. That is the sample must come from the crime scene and cannot just be taken from, for example, a likely suspect's garbage or whisky glass. And law enforcement must have either the suspect's permission or a search warrant (Murphy: 12/23/19). While these are certainly steps in the right direction, they are guidelines only, carry no explicit penalties for noncompliance and still leave the burden of decision and enforcement, and thus the setting of precedent on what are essentially commercial enterprises (Watson: 11/19/19).

A second technology threatening American conventions of privacy is Facial Recognition Technology, often referred to as FRT. A facial recognition system uses biometrics to create a template of facial features gleaned from a video or photograph then compares that template to a database of previously identified faces to find a match. Artificial intelligence is used both to create the template of an unidentified person's face and to search data bases of known persons (Singer: 5/20/17).

The development of FRT has proven useful in a number of ways (Garvie: 10/15/19): it has been used to unlock phones and clear customs. It can also assist doctors in monitoring patients, enable a car to become familiar with the driver's habitual routes and assist stores in taking inventory. It is also useful to law and immigration enforcement in that it can screen footage from surveillance cameras continuously whereas human can only do so attentively for about twenty minutes at a time (Chokshi: 6/13/19).

Of course, there is a potential downside to its use as well. The greatest of these is perhaps the fact that it is often inaccurate. In 2018, the ACLU used Amazon's version of FRT, called Rekognition, to compare photos of all Federal lawmakers with publicly available mugshots. In 5% of the matches members of Congress were erroneously identified with individuals who had been arrested. A disproportionate number of those misidentified were Black or Latinx (Singer: 7/26/18). In 2019 researchers at M.I.T. called on Amazon to stop selling Rekognition because it

was unable to distinguish race and gender accurately. Rekognition mistook women for men 19% of the time and dark skinned women for men 31% of the time (Metz and Singer: 4/3/19).

These problems are complicated by questionable claims by a number of FRT startups that they can detect and identify emotional states and distinguish normal form deviant behavior (Chokshi: 6/13/19), claims that recall the long debunked pseudosciences of phrenology and physiognomy that claimed to be able to assess character and mental capacity from skull shape and facial structure.

This so-called biological essentialism is seeing a resurgence among racial hierarchist like neo-Nazis and white nationalist, but also among academics and facial recognition researchers with politicized agendas. It is evident, for example, in a 2016 paper by Chinese researchers who claimed they could distinguish criminals by measuring features like lip curvature and nose-mouth angle. In 2017 a Stanford University professor claimed to have developed AI gaydar that was 81% accurate. In 2019 the Israeli startup Faceception claimed it could detect terrorists and pedophiles using facial scans and another startup Hirevu claimed to accurately assess videos of job applicants to determine personal stability, conscientiousness and responsibility (Chinoy: 7/10/19).

Another set of concerns about FRT focuses on the sources and scope of the data involved. Police data bases today contain the images of more than half of all Americans, most of whom have no idea that they are there. These images come from a variety of sources including social media, traffic cameras, mug shots and DMV photos. Dozens of databases are being built to improve FRT by giving it's AI more learning examples. Many companies and research institutes share this data with governments, private enterprises and other researchers concentrated on training AI. Share, of course, being a euphemism for sell (Metz: 7/13/19).

By far the greatest sources of such data is that collected and sold by data brokers. This data consists of not only pictures of people's faces but also details about their internet purchasing and browsing histories, records of who they exchange emails with, who they talk to and suggest facts about their incomes, ethnicity and lifestyles. The push to create such databases comes from corporations hoping to influence purchasing behavior but they are also incidentally used by law enforcement and government agencies (Schneier: 1/20/20).

The most recent innovations in FRT is Clearview AI, an app that allows the user to upload a picture of an individual and see publicly posted photos of that individual along with links to where the photos appeared. These photos are supposedly "scraped" from Facebook, YouTube, Venmo and millions of other social media sites, making it the largest FRT database in the world. The app is not currently available to the general public but has reportedly been provided to Homeland Security, the FBI and to hundreds of local law enforcement agencies as well as licensed to a number of private companies as a security tool (Hill: 1/18/20).

The app is popular with law enforcement despite the legal, ethical and operational problems associated with it. Scraping images from social media sites is not specifically illegal but it does not include compensation to the sites' proprietors nor explicit permission from the individuals whose images are being scraped. The app has reportedly been useful in solving crimes like credit card fraud, identity theft, child exploitation and murder, but it also has the potential to facilitate crimes like stalking, blackmail and targeted burglary. Finally, as of this writing, Clearview has never been tested for accuracy or biases as have other FTR's. Like them, it is unlikely to be 100% accurate (Hill: 1/18/20, Pittsburgh Post-Gazette Journal: 2/9/20).

A third area of concern with the use of FRT is its potential politicization. Much of this concern is fueled by the Chinese government's use of this technology not only to enforce social norms like not wearing pajamas in public (Qin: 1/21/20) but also to track members of religious and ethnic minorities and to identify anti-government protesters (Zong: 7/2/19, Mozur: 7/16/19). In the United States, governmental agencies have no clear mandate or explicit legislative authority to use FRT in policing and security control. But neither are there explicit legal checks on its use to identify individuals exercising their Constitutional right to peaceful protest or their right to publicly demonstrate support for controversial opinions and ideas (Friedman and Ferguson: 10/18/19).

In this short paper on genetic genealogy and facial recognition technologies I have attempted to give you some understanding of the threats to privacy that they pose. However, it is important to remember that they are only two pieces of the mass surveillance society that we are in the process of creating. The push to create this surveillance apparatus is coming largely from corporations hoping to predict and influence consumer behavior. But is incidentally and perhaps

increasingly used by law enforcement and government agencies, like ICE, Homeland Security, the FBI and ATF.

Numerous technologies that can be used to identify people without their knowledge or permission already exist or are being developed. They include laser based systems that can detect individual heartbeats or discern individual gaits, cameras so refined that they can read fingerprints and iris scans, smartphones that can broadcast unique traceable numbers, called MAC addresses and scanners that can capture license plate numbers, credit card numbers and phone numbers from a distance.

When correlated with data collected in more pedestrian ways, frequently simply by asking people to volunteer it, the profile of an individual it produces takes on greater substance. This data, purchasing and internet browsing data, email and telephone interlocutor data and data about income, ethnicity, gender and lifestyle helps create a profile that can be used to discriminate as to the kind of advertising one sees, the mail-in offers one gets and the kind of charitable and political support solicited.

Clearly banning this or that technology will not affect this process. In all probability it would simply make data collection, identification and correlation more sophisticated and more secretive. What is needed is rules or laws about when it is or is not alright to identify individuals using mass surveillance technologies, that regulate the data broker industry and that define instances of discrimination as acceptable or unlawful (Schneier: 1/20/20).

Unfortunately, to date lawmakers on both the Federal and State levels have shown little appetite and less ability to formulate such laws. As Charlie Warzel, who has written extensively for the New York Times and other publications on technological issues has argued: “Congress is far from being able to formulate such laws because even though they claim to ‘value internet transparency’, they still do not understand the internet” (Warzel: 6/25/19).

Bibliography

Bala, Nila. Why Are You Publicly Sharing Your Child's DNA Information? New York Times, 1/2/20.

Bambarger, Kenneth; Engleman, Serge; Han, Catherine; Bar On, Amit Elezari; Reyes, Irwin. "Can You Pay For Privacy: Consumer Expectations and the Behavior of Free and Paid Apps." In Van Kiani, Marianne. Privacy Papers 2019, 12/5/19.

Chinoy, Sahil. The Racist History Behind FR. New York Times, 7/10/19.

Choksi, Niraj. How Surveillance Cameras Could Be Weaponized With AI. New York Times, 6/13/19.

Cofone, Ignacio. "Antidiscriminatory Privacy." Southern Methodist University Law Review, 4/19/18.

Frank, Oren. Donate Your Healthcare Data Today. New York Times, 10/2/19.

Friedman, Berry and Gutherie, Andrew Ferguson. Here's a Way Forward on Facial Recognition. New York Times, 10/13/19.

Garvie, Clare. You're in a Police Lineup, Right Now. New York Times, 10/15/19.

Hill, Kashmir. The Secretive Company That Might End Privacy as We Know It. New York Times, 1/18/20.

History.com Editors. This Day In History. A&E Television Networks. 2/26/20.

Igo, Sarah E. The Known Citizen. Harvard University Press, 2018.

Joh, Elizabeth. Want to See My Genes? Get a Warrant. New York Times, 6/11/19.

Kolata, Gina. Your Data Were "Anonymized"? These Scientists Can Still Identify You. New York Times, 7/23/19.

Metz, Cade. Facial Recognition Technology is Growing Stronger, Thanks to Your Face. New York Times, 7/13/19.

_____ and Singer, Natasha. AI Experts Question Amazon's Facial Recognition Technology. New York Times, 4/3/19.

Mozur, Paul. In Hong Kong Protest, Faces Become Weapons. New York Times, 7/16/20.

Murphy, Heather. Sooner or Later Your Cousin's DNA Is Going To Solve a Murder. New York Times, 4/25/19.

_____. Genealogy Sites Have Helped Identify Suspects, Now They've Helped Convict One. New York Times, 7/1/19.

_____. What You're Unwrapping When You Get a DNA Test for Christmas. New York Times, 12/23/19.

_____ and Leveson, Michael. Golden State Killer Suspect Offers to Plead Guilty. New York Times, 3/5/20.

Orenstein, James. I'm a Judge: Here's How Surveillance is Challenging Our Legal System. New York Times, 6/13/19.

Pittsburgh Post Gazette Journal, 2/9/20.

Qin, Amy. Chinese City Uses FR to Shame Pajama Wearers. New York Times, 1/21/20.

Schneier, Bruce. We're Banning FR. We're Missing the Point. New York Times, 1/20/20.

Singer, Natasha. Amazon Faces Investor Pressure Over Facial Recognition. New York Times, 5/20/17.

_____. Amazon's Facial Recognition Technology Wrongly Identifies 28 Lawmakers, ACLU Says. New York Times, 7/26/18.

_____ and Isaac, Mike. Facebook To Pay $550 Million to Settle Facial Recognition Suit. New York Times, 1/29/20.

Watson, Crister. The Third Degree. Fort Wayne Journal Gazette, 11/19/19.

Warzel, Charlie. Congress Wants Data Transparency, But Still Doesn't Understand the Internet, New York Times, 1/25/19.

Zhong, Raymond. China Snares Tourists' Phones in Surveillance Dragnet by Adding Secret App. New York Times, 7/2/19.