AI Ethics Journal

Vol 1, Issue 1
Fall 2020

Data and Privacy Applications

When Artificial Intelligence and Big Data Collide—How Data Aggregation and Predictive Machines Threaten our Privacy and Autonomy

Alexander Alben[1]
[1]*School of Law, University of California, Los Angeles*

DOI: https://doi.org/10.47289/AIEJ20201106

Abstract

Artificial Intelligence and Big Data represent two profound technology trends. Professor Alben's article explores how Big Data feeds AI applications and makes the case that necessity to monitor such applications has become more immediate and consequential to protect our civil discourse and personal autonomy, especially as they are expressed on social media.

Like many of the revolutionary technologies that preceded it, ranging from broadcast radio to atomic power, AI can be used for purposes that benefit human beings and purposes that threaten our very existence. The challenge for the next decade is to make sure that we harness AI with appropriate safeguards and limitations.

With a perspective on previous "revolutionary" technologies, the article explains how personal data became profiled and marketed by data brokers over the past two decades with an emphasis on dangers to privacy rights.

The article observes that it is critical to adopt an approach in the public policy realm that addresses the bias dangers of a technology, while enabling a fair and transparent implementation that allows our society to reap the benefits of adoption. It advocates solutions to improve the technology and adopt the best versions, not cut off development in early stages of the new technology's evolution.

Drawing on the author's work as a state-level Chief Privacy Officer and a high-tech executive, the article concludes with four policy recommendations for curbing the flow of personal information into the Big Data economy: 1. Regulating data brokers; 2. Minimizing data by default; 3. Public Records Reform and 4. Improving personal data hygiene.

History

Received **2 March 2020**

Accepted **1 October 2020**

Published **6 November 2020**

Keywords

Artificial Intelligence, Big Data, Public Records, Data Profiling, Data Hygiene, Data Brokers, Tik Tok, Facebook, Google, IBM, Watson, Thinking Machines, Machine Learning, AI, AI applications, Privacy Rights, Civil Liberties, Data Science

Contact

Alexander Alben

Internet, Law, Media, and Society
UCLA School of Law, University of California, Los Angeles, 85 Charles E Young Dr E, Los Angeles, CA 90095
Email: alben@law.ucla.edu

Acknowledgements

Mr. Alben wishes to deeply thank Guy Bercegeay, a licensed attorney who recently completed an LL.M. program at UCLA School of Law specializing in Media, Entertainment and Technology Law and Policy, for his research and editorial assistance in completing this article. Mr. Alben also wishes to thank Cassandra Frias for her research assistance.

Disclosure of Funding

None

AI Ethics Journal

Introduction

Taking the Long View on New Technology

At the outset of a new decade, we have been promised that Artificial Intelligence (frequently abbreviated as "AI") will solve a host of problems facing our society, ranging from economic disparity to political equity and social injustice. In parallel, we are also told that Artificial Intelligence will create a host of the problems, mapping to the same societal challenges. Both perspectives are correct.

This article examines the relationship between AI and "Big Data." Specifically, it observes that without the fuel of data, the nascent AI industry could not possibly have grown as rapidly or proliferated into the myriad technical implementations we are witnessing today. It also poses the question of whether, as a society, we still have the ability to reign in the flow of Big Data before it truly enables profiling, tracking and prediction of individual human behavior to get out of control, creating more harm than good and threatening personal autonomy. The concluding section offers public policy approaches to curb the flow of the tidal wave of data that is being generated for corporations and data brokers to exploit.

Like many of the revolutionary technologies that preceded it, ranging from broadcast radio to atomic power, AI can be used for purposes that benefit human beings and purposes that threaten our very existence. The challenge for the next decade is to make sure that we harness AI with appropriate safeguards and limitations.

for decades. Anyone who has been a passenger on a jet plane, taken an Uber ride or made a mobile banking deposit has already utilized a form of AI. AI is widely used in e-commerce to identify shopping patterns, prevent fraud and predict future consumer needs.

Despite these widespread uses of AI, a quotient of fear has been introduced into the public discussion, especially in the realm of the threat that AI or even simple "algorithms" pose to civil rights. When *Wired Magazine* ran an article in 2018 citing an ACLU study of Amazon's facial recognition software that erroneously matched 28 of the 435 members of the U.S. Congress with a database of law enforcement mugshots, civil liberties organizations alleged a strong racial bias in the Rekognition software, given that individuals with darker skin tones were twice as likely to be matched with the arrest database with a setting of an 80% confidence level. Amazon noted that other settings in Rekognition could correct for a 95% confidence level in results, yet both privacy advocates and computer scientists chimed in declaring that the software was too likely to make mistakes along racial lines.[1]

Without follow-up, this type of article leaves the indelible impression that some types of AI technologies are inherently prone to error and could result in miscarriages of justice, especially in mistaken identification of suspects. Yet instead of public calls for transparency and better data, many commentators have jumped to the preemptive conclusion that AI for facial recognition should be indefinitely banned. Yet instead of public calls for transparency and better data, many commentators have jumped to the preemptive conclusion that AI for facial recognition should be indefinitely banned. This led

[1] Brian Barrett, *Lawmakers Can't Ignore Facial Recognition's Bias Anymore*, WIRED (July 26, 2018, 4:59 PM), https://www.wired.com/story/amazon-facial-recognition-congress-bias-law-enforcement/.

AI Ethics Journal

to the City of San Francisco to declare a moratorium on the use of AI in 2019.[2] Other jurisdictions have followed suit. A more recent and comprehensive study of the accuracy of facial recognition conducted by the National Institute of Standards and Technology showed that demographic factors do skew AI results[3] and concludes that more caution is needed in the deployment of the technology for use in specific contexts.[4]

The ultimate promise of accurate facial recognition technology, however, must be that it will prevent racial bias by focusing on the characteristics of individuals and not racial traits. Instead of using photography to confirm bias, AI holds out the promise of correctly identifying specific actors as a sorting tool for the implementation of public policy. In the midst of the brouhaha over moratoriums and claims of bias, this larger promise has been lost. This is why it is critical to adopt an approach in the policy realm that addresses the bias dangers of a technology, while enabling a fair and transparent implementation that allows our society to reap the benefits of adoption. We want to look for solutions to improve the technology and adopt the best versions, not cut off development in early stages of its evolution.

The ultimate promise of accurate facial recognition technology, however, must be that it will prevent racial bias by focusing on the characteristics of individuals and not racial traits. Instead of using photography to confirm bias, AI holds out the promise of correctly identifying specific actors as a sorting tool for the implementation of public policy. In the midst of the brouhaha over moratoriums and claims of bias, this larger promise has been lost. This is why it is critical to adopt an approach in the policy realm that addresses the bias dangers of a technology, while enabling a fair and transparent implementation that allows our society to reap the benefits of adoption. We want to look for solutions to improve the technology and adopt the best versions, not cut off development in early stages of its evolution.

Whenever an accident occurs on a self-driving automobile or an algorithm is shown to result in an apparently unfair result, critics contend that all predictive technologies utilizing algorithms are deeply flawed and many contend they should be discontinued pending further study.[5] In short, we are witnessing a frenzy in the technology world over a new set of technologies that we don't completely understand, can't easily define and have reason to both fear and respect.

At the beginning of the computer age in the 1960's, we experienced a similar kind of ambivalence about the coming age where computers would make complex decisions for us, freeing us of painful labor, solving the energy crisis and transforming our economy. Yet today,

[2] Kate Conger et al., *San Francisco Bans Facial Recognition Technology*, N.Y. TIMES (May 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html; *see also* Bruce Schneier, *We're Banning Facial Recognition. We're Missing the Point.*, N.Y. TIMES: THE PRIVACY PROJECT (Jan. 20, 2020), https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html ("Communities across the United States are starting to ban facial recognition technologies. In May [2019], San Francisco banned facial recognition; the neighboring city of Oakland soon followed, as did Somerville and Brookline in Massachusetts (a statewide ban may follow). In December [2019], San Diego suspended a facial recognition program in advance of a new statewide law, which declared it illegal, coming into effect.").

[3] NISTR 8280, Grother, Patrick, Ngan and Hanakoa, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* are intended to inform policymakers and to help software developers better understand the performance of their algorithms. Face recognition technology has inspired public debate in part because of the need to understand the effect of demographics on face recognition algorithms. Full publication available at doi.org/10.6028/NIST.IR.8280.

[4] NIST.gov News: "While it is usually incorrect to make statements across algorithms, we found empirical evidence for the existence of demographic differentials in the majority of the face recognition algorithms we studied," said Patrick Grother, a NIST computer scientist and the report's primary author. "While we do not explore what might cause these differentials, this data will be valuable to policymakers, developers and end users in thinking about the limitations and appropriate use of these algorithms." Reported on NIST.gov, December 19, 2019.

[5] Shepardson, David. Reuters Technology News, November 5, 2019. *In review of fatal Arizona crash, U.S. agency says Uber software had flaws. See also: MIT Media Lab: Algorithmic Justice League Project, available at media.mit.edu.*

AI Ethics Journal

we witness not only the rapid adoption of a new suite of applications utilizing AI, but the marriage of such technology to the explosion of Big Data. It is this combination that makes the issue of accurate AI especially salient and relevant to organizations seeking to deploy AI and to their customers and stakeholders.

As outlined in this article, the data aggregation industry grew in an unregulated fashion and has given rise to a lucrative data broker industry fueled by personal information. While such data profiles are largely used today in commerce, these individual data troves can also be harnessed by governments and other entities. Once created, they are difficult to control. Consequently, lawmakers, scholars and members of the public must become more conscious of the dangers of an unregulated data industry and seriously consider means to regulate the flow of data that will fuel AI applications going forward.

What Are We Talking About?

The Challenge of Defining AI

A good working definition of Artificial Intelligence was floated over ten years ago by Stanford University Computer Science professor Nils Nilsson, a pioneer in the AI field:

> "Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."[6]

Writing in *Forbes Magazine*, Bernard Marr provides more historical perspective on the notion of defining what constitutes AI:

> John McCarthy first coined the term artificial intelligence in 1956 when he invited a group of researchers from a variety of disciplines including language simulation, neuron nets, complexity theory and more to a summer workshop called the Dartmouth Summer Research Project on Artificial Intelligence to discuss what would ultimately become the field of AI. At that time, the researchers came together to clarify and develop the concepts around "thinking machines" which up to this point had been quite divergent.[7]

With this useful working concept of a Thinking Machine, a further refinement of the types of AI is still desirable, given the panoply of technologies that are currently deployed or are on the proverbial drawing board. For the purposes of this article, an Artificial Intelligence program or application will include at least the following elements: 1. The ability to identify data, either through computer language or audio-visual and other "real world" inputs; 2. The ability to store data or seek out data from networked sources; 3. A logic function that allows the program to sort, filter and build hierarchies of data; 4. A machine learning algorithm giving the program the ability to make predictions and to change results based on past experience.

People who have used the United Airlines robotic voice assistant, Ted, have experienced a form of AI that can recognize human language and learn from aggregated chats how to direct consumer queries. PayPal, banks and other financial services use AI programs to detect patterns

[6] NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS, at xiii (2010).
[7] Bernard Marr, *The Key Definitions of Artificial Intelligence (AI) that Explain its Importance,* https://bernardmarr.com/default.asp?contentID=1352

AI Ethics Journal

in commerce that suggest credit card fraud.[8] We are not talking about the types of AI on display in sci-fi movies such as *2001: A Space Odyssey,* where the computer HAL seeks to take over a space mission,[9] although such types of sophisticated programs may become real in our lifetimes.

Given the evolving status of the AI industry, it's interesting how quickly we have come to expect perfection from thinking machines. We seem to live in an environment where every mistake made by a robot or automated vehicle resulting in human injury is widely chronicled and publicized, leading the public to mistrust new technologies that appear to be held to "zero tolerance" standards.[10] Yet another dimension of data is not simply quality, but the sheer number of inputs. The balance of this article focuses on the "bigness" of "big data" and poses whether size itself can result in societal problems when such pools of data are harnessed by AI applications.

How Data Got "Big"

All AI machines benefit from the new world of Big Data, because thinking machines need data in both training and operation. A thinking machine starved of data will not become smart. The proliferation of data types has gone hand in hand with the evolution of computers and the Internet. Embedding cameras in cell phones about 15 years ago has given rise to the creation of more photographs each year than were taken in the previous history of photography.[11] Current estimates for the total number of web sites exceed 1.5 billion, and Google has indexed at least 4.45 billion individual web pages.[12] GPS technology has given rise to tracking data for anyone who keeps their cell phone location data turned on—and even limited tracking when the phone is off.[13]

We have come to tolerate vast aggregations of data as a necessary byproduct of the Internet and connected devices. In the early days of the world wide web, companies worried about the cost of data storage. The advent of cloud services has truly served as a game

---

[8] See generally: John Koetsier, *How Amex Uses AI to Automate 8 Billion Risk Decisions (And Achieve 50% Less Fraud),* Forbes (September 21, 2020), https://www.forbes.com/sites/johnkoetsier/2020/09/21/50-less-fraud-how amex-uses-ai-to-automate-8-billion-risk-decisions/#4a1c5b491a97

[9] For an excellent discussion of four types of machines by an AI researcher, see Arend Hintze, *Understanding the Four Types of Artificial Intelligence*, GOV'T TECH. (Nov. 14, 2016), https://www.govtech.com/computing/Understanding-the-Four-Types-of-Artificial-Intelligence.html.

[10] *See* Daisuke Wakabayashi, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam*, N.Y. TIMES: TECH. (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html ("[A]n autonomous car operated by Uber . . . struck and killed a woman on a street in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology [T]he crash in Tempe will draw attention among the general public to self-driving cars, said Michael Bennett, an associate research professor at Arizona State University 'We've imagined an event like this as a huge inflection point for the technology and the companies advocating for it,' he said. 'They're going to have to do a lot to prove that the technology is safe.'").

[11] *See* Stephen Heyman, *Photos, Photos Everywhere*, N.Y. TIMES (July 29, 2015), https://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html ("The growth in the number of photos taken each year is exponential: It has nearly tripled since 2010 and is projected to grow to 1.3 trillion by 2017. The rapid proliferation of smart phones is mostly to blame."); Amy Hobbs, *[Stats] How Many Photos Have Ever Been Taken?*, FSTOPPERS (Mar. 10, 2012), https://fstoppers.com/other/stats-how-many-photos-have-ever been-taken-5173 (estimating the total number of analogue photographs to be 3.5 trillion).

[12] *See Total Number of Websites*, INTERNET LIVE STATS, https://www.internetlivestats.com/total-number-of-websites/ (last visited Feb. 15, 2020); *How Many Websites Are There Around the World?*, MILL FOR BUS. (Feb. 12, 2020), https://www.millforbusiness.com/how-many-websites-are-there/.

[13] *See* Jennifer Valentino-DeVries et al., *Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret*, N.Y. TIMES (Dec. 10, 2018), https://www.nytimes.com/interactive/2018/12/10/business/location-data privacy-apps.html ("At least 75 companies receive anonymous, precise location data from apps whose users enable location services to get local news and weather or other information . . . . Many location companies say that when phone users enable location services, their data is fair game."); *see also* Ryan Gallagher, *NSA Can Reportedly Track Phones Even When They're Turned Off*, SLATE (July 22, 2013, 4:06 PM), https://slate.com/technology/2013/07/nsa can-reportedly-track-cellphones-even-when-they-re-turned-off.html ("[T]o spy on phones when they are turned off, agencies would usually have to infect the handset with a Trojan that would force it to continue emitting a signal if the phone is in standby mode, unless the battery is removed. In most cases, when you turn your phone off . . . it will stop communicating with nearby cell towers and can be traced only to the location it was in when it was powered down.").

AI Ethics Journal

changer in making Big Data even bigger.[14] Cheaper and cheaper storage has driven the cost of retaining data to zero or below, creating a paradox that it is now cheaper for most firms to keep data than to delete it. If true, this will only accelerate data proliferation.[15]

Artificial Intelligence technologies have existed for decades, yet only in the past ten to fifteen years have they been married to pools of large data, enabling them to accomplish both useful and invasive tasks.

In 2011, IBM's Watson computer made headlines by defeating two accomplished Jeopardy Champions, Ken Jennings and Brad Rutter, in a three-game match.[16] The playful "Smarter Planet" logo that viewers saw on television masked ten racks of IBM Power 750 servers sitting in a separate room. When Watson cogitated host Alex Trebeck's questions, wavy green lines animated his "face."

Unlike many AI applications that can process natural language, Watson did not actually listen to Trebeck, but received his inputs via text messages that translated the host's verbal questions. However, like his two human contestants, Watson had to frame his questions in terms of answers and had to "buzz in" to gain priority to answer correctly. He did so with stunning accuracy, despite the fact that he had no Internet connection as many viewers assumed. In fact, Watson had been fed over 200 million pages of data, ranging from sports to entertainment trivia. By besting Jennings and Rutter by over $50,000 Jeopardy dollars, Watson claimed the one-million-dollar tournament prize.

Watson's AI architecture had taken IBM scientists over three years and thousands of practice rounds to develop. According to *TechRepublic:*

> IBM developed DeepQA, a massively parallel software architecture that examined natural language content in both the clues set by Jeopardy and in Watson's own stored data, along with looking into the structured information it holds. The component-based system, built on a series of pluggable components for searching and weighting information, took about 20 researchers three years to reach a level where it could tackle a quiz show performance and come out looking better than its human opponents.[17]

Watson's victory sparked widespread interest in AI technologies, even though Watson itself would be further developed by IBM for commercial applications. Health insurer Wellpoint and The Memorial Sloan Kettering Medical Center began utilizing Watson for health care problem solving by 2012.[18] To do so, Watson not only was put on-line, but had to learn to properly ingest medical taxonomies and two million pages of medical data from over 600,000 sources. The value of using this type of AI to sort through presentations of facts and make diagnoses had not been lost on the oncologists at Sloan-Kettering. While used only as a back-up to human diagnosticians, Watson came to make accurate diagnoses and was able to

---

[14] *See* Mike Chan, *Big Data in the Cloud: Why Cloud Computing Is the Answer to Your Big Data Initiatives*, THORN TECHS. (Sept. 10, 2018), https://www.thorntech.com/2018/09/big-data-in-the-cloud/ (discussing several key advantages of combining "big data analytics and cloud computing").

[15] *See* Michael Fertik, *Why Your Data Will Never Be Deleted*, Forbes (June 9, 2015, 10:14 AM), https://www.forbes.com/sites/michaelfertik/2015/06/09/why-your-data-will-never-be-deleted/#40fb590a2371.

[16] See generally: John Markoff, *Computer Wins on 'Jeaopardy!': Trivial, It is not.* N.Y. TIMES (February 16, 2011), https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

[17] Jo Best, *IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next,* Tech Republic (September 9, 2013), https://www.techrepublic.com/article/ibm-watson-the-inside story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/.

[18] See generally: IBM and Wellpoint put #ibmwatson to work in Healthcare, IBM, https://www.ibm.com/support/pages/ibm-and-wellpoint-put-ibmwatson-work-healthcare.

incorporate patient history and DNA testing in devising its predictions of what course of treatment would be appropriate for cancer patients.

Watson has also been adapted for use in the finance industry and for customer service applications. It has already been tested by some banks that use Watson to recommend financial services to customers. Sites developed by law firms have trained Watson's Natural Language functions to help answer basic legal questions.[19] The platform is even finding its way into retail settings, attempting to influence the course of consumer purchasing decisions.[20] Like all powerful technologies developed before it, AI will also be harnessed for entertainment and less than life-or-death human pursuits. Edge Up Sports, a Fantasy Football start-up, for example, has employed Watson to give its Fantasy Fans better data recommendations than they might otherwise develop by consumer sports stats on their own.[21]

We can surely expect AI platforms, such as Watson, to be trained to tackle challenges in health care, finance, customer service and myriad verticals over the next decade. With human supervision and back-up, these powerful programs might improve diagnostic accuracy and speed up critical processes that make a difference in human lives. At least that's the promise of AI, yet the collision of the AI and Big Data train have already led to the challenges of technologies quickly getting out of control, akin to the metaphor of opening Pandora's Box.

Big Data Meets Social Media Platforms

One of the first pools of data harnessed by AI algorithms has been the personal information that billions of users of Twitter, Google, Amazon, Microsoft and other leading Tech platforms have generated in pursuit of communications and social interaction. The irony that users don't adequately value their own data has not been lost on a growing number of economists and legal scholars, who have suggested that the fundamental business models of these social networks is flawed from a consumer perspective. The defining paradox of the age of social media may turn out to be that while each user is willing to trade their most personal data for free software, the value of such data in aggregate is exploited by some of the most profitable enterprises that the world has ever seen.

At the moment, TikTok stands out as a prominent example of an application that utilizes algorithms and vast troves of user data to create a compelling entertainment product. Owned by a Chinese parent company, ByteDance, the app has raised the ire of the Trump Administration due to potential data sharing of user information with the Chinese government. Tik Tok's global reach is difficult to dispute. As of July 2020, the song and dance-centered social application had over 689 million users across the planet and has been downloaded over two billion times. In the U.S., it has reached approximately 100 million monthly users, an increase of over 800% in two years.[22]

---

[19] *How Watson helps lawyers find answers in legal research: ROSS Intelligence takes Watson to law school*, MEDIUM.COM (Jan. 4, 2017), https://medium.com/cognitivebusiness/how-watson-helps-lawyers-find-answers-in legal-research-672ea028dfb8.

[20] ZDNet, "IBM Watson: What Are Companies Using it For," Conner Forrest, September 1, 2015.

[21] Edge Up Sports, LLC: Using Cognitive technology to help fantasy football "owners" make better roster decisions. IBM, https://www.ibm.com/case-studies/a787535m28346z55.

[22] CNBC, "TikTok Reveals Detailed User Numbers For the First Time," August 24, 2020. This article includes the following data:
TikTok revealed specific U.S. and global growth milestones for the first time in a lawsuit against the U.S. government.
TikTok has about 100 million monthly active U.S. users, up nearly 800% percent from Jan. 2018. TikTok said it has been downloaded about 2 billion times globally.
TikTok said it has about 50 million daily active U.S. users
Here's the breakdown of TikTok's U.S. user growth: January 2018: 11,262,970 U.S. monthly active users (MAUs), February 2019: 26,739,143, October 2019: 39,897,768
June 2020: 91,937,040, August 2020: More than 100 million based on quarterly usage globally, TikTok has experienced similar surges in users. The company said it had about 55 million global users by Jan. 2018. That number ballooned to more than 271 million by Dec. 2018 and 507 million by Dec. 2019. This month, TikTok surpassed 2 billion global downloads and reported nearly 700 million monthly active users in July.

AI Ethics Journal

While the Trump administration did not cite any specific instance of user data transferred to Chinese authorities, the company has been criticized for violating Google's Android Platform privacy policy regarding the capture of user device MAC addresses. In August of 2020, the Wall Street Journal outlined how TikTok violated the Android policy:

> TikTok skirted a privacy safeguard in Google's Android operating system to collect unique identifiers from millions of mobile devices, data that allows the app to track users online without allowing them to opt out, a Wall Street Journal analysis has found. The tactic, which experts in mobile-phone security said was concealed through an unusual added layer of encryption, appears to have violated Google policies limiting how apps track people and wasn't disclosed to TikTok users. TikTok ended the practice in November, the Journal's testing showed.[23]

Clearly, the potential exists for consumer data to find its way from a TikTok consumer's device to a parent company and then to entities that are not identified in TikTok's end user license agreement. Highly popular with children and teens, the specter of unauthorized collection and transfers of user data also raises serious questions relating to children's privacy, compliance with COPPA and similar legislation in the European Union and other jurisdictions. The fact that user-generated videos also feature images of TikTok users, also creates possibilities of utilizing facial recognition and other identifiers, so that a person appearing in an apparently harmless homemade karaoke may in fact by spotted and tracked for nefarious purposes.

Like any software product, TikTok is powered by algorithms, which lie at the heart of the struggle between the perceived interests of the American government and the autonomous operation of ByteDance, TikTok's owner. Based on a blog post by the company, *Wired Magazine* reported a few details of how the "For You" function works, determining which videos a user will see in their TikTok app:

> When a video is uploaded to TikTok, the For You algorithm shows it first to a small subset of users. These people may or may not follow the creator already, but TikTok has determined they may be more likely to engage with the video, based on their past behavior. If they respond favorably—say, by sharing the video or watching it in full— TikTok then shows it to more people who it thinks share similar interests. That same process then repeats itself, and if this positive feedback loop happens enough times, the video can go viral. But if the initial group of guinea pigs don't signal they enjoyed the content, it's shown to fewer users, limiting its potential reach.[24]

While this type of affinity algorithm has been in use for decades, suitors for the company apparently consider it to be of great value and have insisted that ownership of the algorithm be part of the resolution of an acquisition of the company. Similarly, Facebook has created controversy and criticism by weighting user "News Feeds" with factors that stimulate polarized views on political issues, although we lack a definitive scale of what type of speech might be purely factual or neutral.

[23] *Kevin Poulsen and* Robert McMillan, *TikTok Tracked User Data Using Tactic Banned by Google,* Wall Street Journal, Online.; Updated Aug. 11, 2020 4:58 pm ET
[24] Wired Magazine, TikTok Finally Explains How the 'For You' Algorithm Works, June 18, 2020.

Nevertheless, critics of Facebook decry the "Filter Bubbles" that determine exposure of content to its user base and argue that its AI technology has disrupted political discourse and caused societal harm:

> Where Facebook asserts that users control their experience by picking the friends and sources that populate their News Feed, in reality an artificial intelligence, algorithms, and menus created by Facebook engineers control every aspect of that experience.[25]

Facebook can draw upon the user profiles and activity generated each second by Instagram and Facebook and WhatsApp platforms, allowing it to slice and dice data in ways that cater to user preferences and are driven by user behavior on the site. In its first five years, when activity primarily revolved around the social experiences of its college age members, Facebook attracted little criticism. Once the "Like" button was added in February of 2009, the company gained a powerful input to understand user behavior and serve content more closely tailored to the needs of its advertisers.[26] "Likes" fueled the growth of the platform from 400 million to over two billion global users, illustrating how a smart company can harness the potential of Big Data.

Facebook can draw upon the user profiles and activity generated each second by Instagram and Facebook and WhatsApp platforms, allowing it to slice and dice data in ways that cater to user preferences and are driven by user behavior on the site. In its first five years, when activity primarily revolved around the social experiences of its college age members, Facebook attracted little criticism. Once the "Like" button was added in February tailored to the needs of its advertisers.[26] "Likes" fueled the growth of the platform from 400 million to over two billion global users, illustrating how a smart company can harness the potential of Big Data.

Information that a user provides Facebook isn't limited to elements such as likes, posts and photos, but can include the location metadata inside photos, and even what is seen through the camera in its apps. Facebook uses a person's address book, call log or SMS log to suggest people that the user may know. The company can collect a user's phone number and additional information from other people uploading their contacts.

Whenever possible, Facebook logs each individual's phone's battery level, signal strength, even available storage.[27] On a computer, Facebook logs a user's browser type and its plugins. It also tracks whether a window is in the foreground or background, and the movements of a mouse. While Facebook can obtain location data when provided access to GPS, the company doesn't stop tracking an individual's location when they turn off location services. It also tracks location from other data points, including IP addresses and nearby Wi-Fi access points and cell towers.[28]

Facebook also gathers information about other devices that are nearby or "on your network." The policy says it is to make it easier, for instance, to stream video from your phone to your TV.[29] Because Facebook provides proper

[25] McNamee, Roger, "Zucked—Waking Up to the Facebook Catastrophe," pp. 90-91.
[26] Fast Company, "How Facebook's 'Like' button Hijacked our Attention and broke the 2010s," by Christopher Zara, December 18, 2019.
[27] Katherine Bindley & Wilson Rothman, *Facebook Has a New Data Policy—Here's the Short Version*, WALL ST. J. (Apr. 20, 2018, 9:29 AM ET), https://www.wsj.com/articles/facebook-has-a-new-data-policyheres-the-short-version 1524230950.
[28] Kashmir Hill, *Turning Off Facebook Location Tracking Doesn't Stop It from Tracking Your Location*, GIZMODO (Dec. 12, 2018, 12:20 PM), https://gizmodo.com/turning-off-facebook-location-tracking-doesnt-stop-it-f-1831149148 ("'Facebook does not use WiFi data to determine your location for ads if you have Location Services turned off,' said a Facebook spokesperson by email. 'We do use IP and other information such as check-ins and current city from your profile.'").
[29] Jake Kanter *Facebook Is Tracking You in Ways You Never Knew – Here's the Crazy Amount of Data It Sucks up*, BUS. INSIDER (June 12, 2018, 2:12 AM), https://www.businessinsider.com/facebook-reveals-all-the-way-it-tracks-user behaviour-2018-6.

notice of these practices, all of this data gathering is perfectly legal, and the company goes on to detail its data sharing practices. The FTC fined Facebook a record $5 billion in 2019 for failure to comply with a previously signed consent decree regarding its data sharing practice and agreed to internal structural changes relating to both data governance and the security of user personal information.[30] Separately, Facebook agreed to a $100 million settlement with the Securities and Exchange Commission relating to misuses of data stemming from its public failure to disclose its failure to correct the damage caused by the data leak in the Cambridge Analytica scandal of 2015.[31]

An individual human could not track this amount of data about themselves, giving rise to the phenomenon that large digital platforms "know more about us than we know about ourselves." One might counter that a large social media platform knowing a trivial fact about a person, such as the battery level of a phone at any given moment, is inconsequential.

In fairness, the acquisition of each such minor data point is arguably quite harmless in and of itself. However, large platforms acquire multiple and diverse data points for the purpose of building a more comprehensive and nuanced picture about the life of the individual subject, combining such data on a scale that has not been previously attained and with potentially dire consequences. Data collected with a consumer's knowledge for a specified purpose lives on permanently for other uses and might even be transferred to third parties without that person's consent.

This "problem of big data" can also arise on non-social applications, which users trust with their personal information. Google maps an individual's behavior by location over the years based on your access to Google and Android services,[32] such as Gmail or apps on an individual's smart phone. Here's a snapshot of the Author's travels over a one-year period, collected by Google without advance notification and subject to deletion only if a Google application user discovers the link where the map resides. At minimum, a policy respecting user privacy would make such collection of data an "opt in" experience. Certainly, such personal data would fall under California's new California Consumer Privacy Act, which went into effect on January 1, 2020.

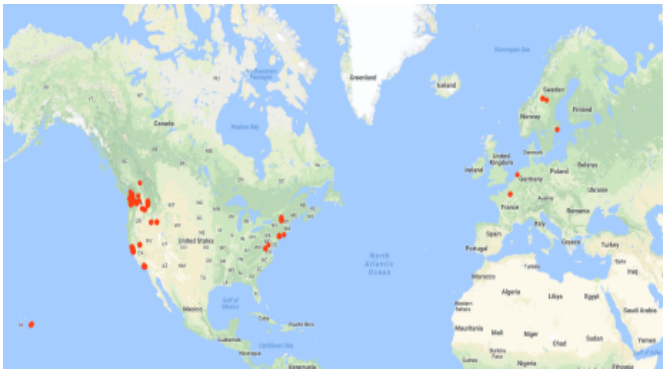
---

[30] In particular, Facebook must "expand its privacy protections across Facebook itself, as well as on Instagram and WhatsApp. It must also adopt a corporate system of checks and balances to remain compliant, . . . [and] maintain a data security program, which includes protections of information such as users' phone numbers." Mark Zuckerberg's oversight in privacy and security matters also has been diminished under the FTC settlement. Specifically, Facebook must "create a new privacy committee with independent board members who cannot be removed without a two
thirds shareholder vote. Zuckerberg and designated compliance officers each must submit individual quarterly compliance reports to the FTC." Finally, "a third-party assessor will monitor Facebook's privacy-related decisions going forward." Mike Snider & Edward C. Baig, *Facebook Fine $5 Billion by FTC, Must Update and Adopt New Privacy, Security Measures*, USA TODAY (July 24, 2019, 8:54 AM ET),
https://www.usatoday.com/story/tech/news/2019/07/24/facebook-pay-record-5-billion-fine-u-s-privacy violations/1812499001/ ("The $5-billion FTC fine is nearly 20 times greater than the largest privacy or data security penalty that has ever been assessed worldwide and is one of the largest imposed by the U.S. government for any violation . . . . The commission approved the settlement with a 3-2 vote, with the dissenting commissioners wanting tougher action taken against Zuckerberg.")

[31] *Id.* ("[The SEC complaint alleged that] Cambridge Analytica paid an academic researcher to 'collect and transfer data from Facebook to create personality scores for approximately 30 million Americans' and that Facebook discovered this misuse in 2015 but failed to correctly disclose it for more than two years."). [32] The Author's Google Maps Timeline for a one-year period, provided by Alex Alben. For more on this point, see: *Google Is Tracking Your Location—Even Without Your Permission, Report Says,* Fortune Magazine, August 13, 2018; https://fortune.com/2018/08/13/google-tracking-locations-without-permission/

[32] The Author's Google Maps Timeline for a one-year period, provided by Alex Alben. For more on this point, see: *Google Is Tracking Your Location—Even Without Your Permission, Report Says,* Fortune Magazine, August 13, 2018; https://fortune.com/2018/08/13/google-tracking-locations-without-permission/

AI Ethics Journal



Amazon not only knows a person's purchase history, but how often they leave your shopping cart open. The public may be aware of these practices, but very few consumers object or even dial down the privacy settings made available by these companies.[33]

This level of comprehensive data gathering is akin to the "mosaic theory" of Fourth Amendment surveillance law, which holds that an isolated photo or video might not constitute an intrusion or a "search," yet stringing together multiple photos or videos could most likely yield a detailed account of an individual's pattern, habits and life.[34]

The harm created by broad data profiling is not the collection of random data points about a person, but the aggregation of such points to paint a complex profile of an individual and her history and predilections.

This is the danger of Big Data when harnessed to AI programs such as machine learning.

How Our Data Became Brokered

Social media platforms acquire even more information about their users from data brokers such as Acxiom and Oracle. These firms specialize in the collection of data and monetize it through sales to marketers, to both traditional firms and online. One might observe that the White Pages published by local telephone operators was an early form of data collection and in the pre-digital era, records were regularly compiled by the county clerk to measure births, deaths, mortgages and property sales. Yet before the advent of widely available search algorithms, the collection and analysis of these written records required considerable investment of time and labor. With the advent of Big Data, the data broker industry struck the mother lode. Disparate databases could be harnessed together. Different data types could be sorted, analyzed and recombined. Individuals could be tracked across hundreds and even thousands of data repositories. Acxiom compiles data on individuals to track their: religion, health interests, alcohol and tobacco consumption, banking relationships, social media usage, medical insurance, size and type of home, family size and likelihood of having another baby, loans, income, personal net worth, relationship status, media consumption, political views and, of course, age, gender, education and employment.[35]

---

[33] *See* Janko Roettgers, *Facebook Exec Doesn't Expect Privacy Backlash to Impact Revenue*, VARIETY (Apr. 13, 2018, 9:39 AM PT), https://variety.com/2018/digital/news/facebook-exec-privacy-backlash-revenue-1202752652/ ("[According to Facebook's vice president of global marketing solutions, Carolyn Emerson,] the company hasn't seen many people change their privacy settings on the service. 'People are going in checking it out, for sure,' she said. [But] that curiosity doesn't necessarily result in any changes, at least not for now. 'We are not seeing a surge in any changing in consumer behavior,' she said."); *see also The State of Privacy in Post-Snowden America*, PEW RESEARCH CTR. (Sept. 21, 2016), https://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/ [hereinafter *The State of Privacy*] ("Even after news broke about the NSA surveillance programs, few Americans took sophisticated steps to protect their data, and many were unaware of robust actions they could take to hide their online activities.");

[34] Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 320 (2012) ("The mosaic theory requires courts to apply the Fourth Amendment search doctrine to government conduct as a collective whole rather than in isolated steps. Instead of asking if a particular act is a search, the mosaic theory asks whether a series of acts that are not searches in isolation amount to a search when considered as a group. The mosaic theory is therefore premised on aggregation: it considers whether a set of nonsearches [sic] aggregated together amount to a search because their collection and subsequent analysis creates a revealing mosaic.").

[35] Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, N.Y. TIMES (June 16, 2012), https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html ("[Acxiom] is integrating what it knows about our offline, online and even mobile selves, creating in-depth behavior portraits in pixilated detail. Its executives have called this approach a '360-degree view' on consumers.").

AI Ethics Journal

Other leading data brokers include Oracle, Experian, Trans Union, Lifelock, Equifax, Moody's and Thomson-Reuters. Each systematically gathers personal information from public sources, purchases other personal data from private sources and monetizes their different user profiles to meet the needs of data-hungry customers. To date, over 400 firms have identified themselves under the data broker provisions of California's new privacy law. When Vermont passed the first data broker registry law in 2018, over 120 data brokers eventually signed up and paid the $100 registration fee.[36] While it is difficult to arrive at a single definition of "data" or "information" broker, it is thought that the global industry rakes in between $200 and $300 billion annually.[37] Thus, personal data has become the fuel for a secondary market that simply trades in a widely available resource—our personal data.[38]

Four Solutions to the Big Data Problem

The operation of the data broker industry is perfectly legal in the United States and data brokers have had no obligation until recently to even tell an individual what data it has collected about them. (In January of 2020, California's new privacy law established a right to request access to personal data from certain firms, including data brokers.) AI sits atop the iceberg of Big Data. Whether we are talking about image recognition or other forms of machine learning, data-hungry AI machines thrive on ingesting data to form and perfect their ability to navigate the "real world" and to solve problems they are trained to solve. One could even go so far as to posit that the AI industry could not exist in a meaningful way without huge depositories of data on which to be trained and produce results. In the perfect storm of this new decade, the AI industry doesn't need to worry about finding data, just as a 6th grader doesn't have to worry about finding an online source for a book report.

Can Big Data Be Controlled?

Like King Canute attempting to hold back the advancing tide, proposals to curb Big Data might strike most people in the technology industry as whimsical or quixotic. There are practical measures, however, that can influence the course of data creation and data retention which we need to more fully explore in order to ascertain when they are effective in giving users more control over their personal information:

1. Regulation of Data Brokers

With regard to the regulation of data brokers, efforts are underfoot in several states to identify data brokers and create more consumer transparency around their practices. Vermont became the first state to pass a data broker registration statute in 2018. The new Vermont law defines a "data broker" as a business that collects and sells

[36] *See* VT. OFFICE OF THE ATTORNEY GEN., GUIDANCE ON VERMONT'S ACT 171 OF 2018 DATA BROKER REGULATIONv6 (2018), https://ago.vermont.gov/wp-content/uploads/2018/12/2018-12-11-VT-Data-Broker-Regulation-Guidance.pdf; *see also* Steven Melendez, *A Landmark Vermont Law Nudges over 120 Data Brokers out of the Shadows*, FAST COMPANY (Mar. 2, 2019), https://www.fastcompany.com/90302036/over-120-data-brokers-inch-out-of-the-shadows-under landmark-vermont-law ("The law also requires companies to spell out whether there's any way for consumers to opt out of their data collections, to specify whether they restrict who can buy their data, and to indicate whether they've had any data breaches within the past year.").

[37] Jason Morris & Ed Lavandera, *Why Big Companies Buy, Sell Your Data*, CNN: BUSINESS, https://www.cnn.com/2012/08/23/tech/web/big-data-acxiom/index.html (last updated Aug. 23, 2012, 3:52 PM ET) ("Data is now a $300 billion-a-year industry and employs 3 million people in the United States alone . . . .").

[38] For a brief discussion of how the General Data Protection Regulation ("GDPR") effectively prohibits the formation of a personal-data secondary market in Europe, compare Chiara Rustici, *Personal Data and the Next Subprime Crisis*, FORBES (July 24, 2018, 2:17 PM), https://www.forbes.com/sites/chiararustici/2018/07/24/personal-data-and-the next-subprime-crisis/#6d60080170aa (suggesting that "there will never be a deep and liquid personal data ***secondary*** market" because the GDPR: (1) prohibits re-purposing lawfully-collected personal data; (2) absolutely bars utilizing such data "past an agreed time frame"; and (3) heavily restricts "any onward-transfer" of such data—i.e., "personal data may behave as an asset, . . . but it will never behave as a commodity [in Europe]") (emphasis in original).

personal information from consumers with whom the broker has no direct relationship. Thus, the Vermont law begins to address "third party" data mining (that is, data mining by companies that have no direct relationship with consumers).[39]

In the wake of its ground-breaking new privacy law, the California Legislature also passed a bill requiring data broker registration at the end of 2019.[40] Other states are considering data broker registration. When coupled with the CCPA's requirement that a consumer can request access to and deletion of data from a company that garners more than half of its revenue from the sale of personal information, California has now jumped the queue in terms of transparency.

Transparency is the fundamental principle in this regime. Once people are aware of the actual practices of data brokers and how those practices impact their personal lives, they may act to curtail certain types of data sharing and surely will become more sympathetic to legislative efforts, perhaps even a national law, to restrict data broker practices that freely trade their personal data without the need to seek their consent.[41] Third party data sharing provides the oxygen for the fire of the data industry, further enabling applications that make predictions about user behavior.

2. Deleting Content by Default

Companies love to keep data, yet if our society is to bring Big Data under control, the collection and retention of data should be purposeful and driven by either an identified company need—which could be monetization—or consumer benefit.

Companies keep data because they can and the machines are set by default to chronicle records of time spent on site, pages visited, pages hovered over and other metrics. While such data can be useful for analysis of consumer behavior, that does not justify the retention of all data from all users.

Keeping all data for indefinite periods of time poses hazards for companies, opens them up for claims of unjustified tracking and contributes to the big data problem. Leading private sector actors with sophisticated engineer programs clearly are capable of doing better and doing more with limited data sets, yet their feet have never really been held to the fire, either by regulators or the public at large. With the growing recognition of the lead of data to unintended places and growing legislative calls for data scrutiny, now would be a good time to begin by shifting the default setting to the deletion of data and to conscious and transparent justification for data retention.

3. Public Records Reform

In several states, public records acts have become the third rail of politics and unintentionally have contributed to the flow of data from public records to data brokers and other

[39] Adam Schwartz, *Vermont's New Data Privacy Law*, ELECTRONIC FRONTIER FOUND. (Sept. 27, 2018), https://www.eff.org/deeplinks/2018/09/vermonts-new-data-privacy-law ("But it does not address "first-party" data mining . . .. For example, the Vermont law does not cover a social media platform like Facebook, or a retailer like Walmart, when those companies gather information about how consumers interact with their own websites.").

[40] *CCPA and California's New Registration Requirement*, NAT'L L. REV. (Sept. 16, 2019), https://www.natlawreview.com/article/ccpa-and-california-s-new-registration-requirement. The California law defines "data broker" in the same manner as the Vermont law, and it exempts credit reporting agencies (covered by the Fair Credit Reporting Act), financial institutions (covered by the Gramm-Leach-Bliley Act) and covered entities under HIPAA. *Id.*

[41] *See Public Opinion on Privacy*, EPIC.ORG, https://epic.org/privacy/survey/ (citing Sam Sabin, *Most Voters Say Congress Should Make Privacy Legislation a Priority Next Year*, MORNING CONSULT (Dec. 18, 2019, 12:01 AM ET), https://morningconsult.com/2019/12/18/most-voters-say-congress-should-make-privacy-legislation-a-priority next-year/) ("A new poll of registered voters found that 79% of Americans believe that Congress should enact privacy legislation and 65% of voters said data privacy is 'one of the biggest issues our society faces.'").

AI Ethics Journal

other commercial actors.[42] Originated in the 1970's, most public records acts seek to spread "sunshine" in the workings of government by mandating that administrative records be retained for set periods of time and be made available to the public at little or no cost.[43] Newspapers and broadcasters rely on public disclosure to uncover news about the workings of state and local government and to conduct investigations relating to individuals. No one questions this legitimate use of the public disclosure system. Yet, as described below, other actors have utilized the mountain of public data for their own uses with no regard to the broader public interest. Unfortunately, efforts at public records reform to address specific abuses and abusers, are usually met with criticism that such reform aims to limit "the public's right to know." With the evolution of Big Data, this type of argument has become increasingly divorced
from the reality of how public records are actually consumed.

In the era of file cabinets and clerks, the flow of public records relating to real estate transactions, births and deaths and criminal records grew at a reasonable pace. However, as in the wider industry, the advent of new data types such as audio, video, GPS and social media communications, vastly exploded the scope of public records, with most courts holding that a new data type qualified as such.[44] Further, the evolution of online search made these records readily available, both to interested members of the public and to commercial entities seeking to scoop up data on a routine basis. Much of the data industry, in fact, relies on "scraping" public data bases for millions of individual records.[45] The brokers then combine this public information with personal data gleaned from other sources, creating rich and more economically valuable profiles.

Commentators have pointed out that the long tail of personal data often results in distorted profiles for individuals, especially people who have gone through the criminal justice system.[46] Further, the compilation of

---

[42] *See* Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1967, 1972, 1988 (2015) ("Executive agencies frequently release government information under sunshine laws such as the Freedom of Information Act (FOIA), which requires disclosures in response to public records requests provided that no law prohibits the release . . . . Private companies, such as data brokers and app developers, are compiling information from public records, combining it with information from other sources, and repackaging the combined information as new products or services.") (citations omitted); Kirsten Martin & Helen Nissenbaum, *Privacy Interests in Public Records: An Empirical Investigation*, 31 HARV. J.L. & TECH. 111, 120 (2017) ("[C]ommercial stakeholders, such as data brokers, enjoy greater efficiencies in their bulk collection of data from public records, from which they extract knowledge that is attractive to other stakeholders in various sectors . . . .") (citations omitted).

[43] *See, e.g.*, WASH. REV. CODE § 42.56.100 (West, Westlaw through ch. 2 of 2020 Reg. Sess.) ("Agencies shall adopt and enforce reasonable rules and regulations . . . to provide full public access to public records, to protect public records from damage or disorganization . . . ."); *see generally* LOUIS D. BRANDEIS, OTHER PEOPLE'S MONEY: AND HOW THE BANKERS USE IT 91 (Martino Fine Books 2009) (1914) ("Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.").

[44] *See* Nissen v. Pierce Cty., 357 P.3d 45, 55-56 (Wash. 2015) (en banc) (holding that content of work-related text messages sent and received by a county prosecutor on his private cell phone in his official capacity were "public records" under the Public Records Act); West v. Puyallup, 410 P.3d 1197, 1201 (Wash. Ct. App. 2018) (noting that a public official's posts on a personal social media page can constitute an agency's public records subject to disclosure under the Public Records Act if the posts relate to the conduct of government and are prepared within a public official's scope of employment or official capacity).

[45] *See* LEXISNEXIS, TEN COMPELLING REASONS TO RELY ON LEXISNEXIS® PUBLIC RECORDS AS YOU RESEARCH PEOPLE, BUSINESSES AND LOCATIONS 1 (2012), http://www.lexisnexis.com/pdf/Ten%20Reasons_Corp_Gov_FINAL.pdf ("Access more than 36 billion public records—one of the world's largest online collections. [C]ast the broadest net possible to capture the most current and complete picture of a person, business or location. LexisNexis Public Records also offers more types of records, including email, health-care provider sanctions, employment locator[] and others . . . ."); *see also* Ms. Smith, *Cha-Ching of Scraping: Data Brokers Digging up & Selling Your Digital Dirt*, CSO: PRIVACY &SECURITY FANATIC (Oct. 12, 2010, 1:25 PM PST), https://www.csoonline.com/article/2227434/cha-ching-of-scraping--data-brokers
digging-up---selling-your-digital-dirt.html ("[S]ite scraping happens all the time, ranging from free do-it-yourself scraping software to screen-scrapers that charge between $1,500 and $10,000 for most jobs. The website PatientsLikeMe.com discovered media-research Nielsen Co. was scraping all messages off the private online forum, messages that were supposed to be viewable only by members who have agreed not to scrape, and not by intruders such as Nielsen.") (citations & internal quotation marks omitted)

[46] *See* Colleen V. Chien & Jason Tashea, *Better Data and Smarter Data Policy for a Smarter Criminal Justice System System*, THE ETHICAL MACHINE (Dec. 10, 2018), https://ai.shorensteincenter.org/ideas/2018/12/10/better data-and-smarter-data-policy-for-a-smarter-criminal-justice-system-system ("Across the country, courts are using profile-based risk assessment tools to make decisions about pretrial detention. The tools are built on aggregated data about past defendants to identify factors that correlate with committing a subsequent crime or missing a trial date. They are used to score individuals and predict if pretrial incarceration is necessary because these tools are built on

public records has enabled bad actors to utilize the public records act to harass victims, such as estranged spouses or people whom they bear a grudge against.[47]

This was not the intent of the 1970's sunshine era laws, yet the proliferation of data created a wave of information that has swamped the resources of states and local government.

A modest and targeted reform of the public records act could achieve the goals of limiting the flow of data while preserving the rights of journalists and members of the public. First, a public records requester should have to state a reason for the request, such as their personal need to find information about another individual. The records keeper can then have some basis for evaluating a request and flag suspicious requests. Second, data brokers should have to purchase data from a government according to an approved agreement. Many of these exist in the realm of departments of motor vehicles for vehicle data and such agreements should be duplicated across other state functions, such as taxation and business records. Third, states should increase penalties for those seeking commercial use of public records and pro-actively try to stop such requests when they are made. For example, a request for "all addresses of homeowners in a water district," should give rise to suspicion that the requestor is seeking the information for a commercial purpose. A request to see the recorded video of a woman's movement in and out of a public building should be treated with suspicion, especially in the context of a sexual harassment inquiry. At present, most state laws place the burden on the government to establish that the request is suspect.

Many of the new data types created in public and private sectors are "transient." For example, GPS location data from a state vehicle is captured and recorded, but not necessarily kept by the wireless carrier or intermediary for longer than a day or two. However, strict reading of public records statutes calls for the retention of such data. One logical reform would be to more carefully define and examine so-called "transient" data. Limiting the mandatory collection of transient data will protect the privacy interests of such civil servants, while also narrowing the funnel of data that has contributed to the overload and increasing misuse of public data by private actors with motives that do not meet the legislative goals of public transparency.

Finally, retention periods for public data should be reviewed. Most of these retention periods were set in a pre-digital pre-search era, where there was a greater justification to keep data around longer. If data retention periods for public records can be shortened across the board, those seeking timely news and information will be favored As it stands, the system rewards massive scoops of public data that end up in

historical data, they run a real risk of reinforcing the past practices that have led to mass incarceration, like the over incarceration of poor and minority people."); *see also* Colleen V. Chien & Clarence Wardell III, *Make the First Step Act a Smarter Step by Opening the Risk Assessment Black Box*, THE HILL:CRIM.JUST.(Dec. 16, 2018, 8:00 AM EST), https://thehill.com/opinion/criminal-justice/421552-make-the-first-step-act-a-smarter-step-by
opening-the-risk?amp ("[T]he reliance on factors like a person's history, educational background, and other demographic factors to classify them risks exacerbating and further embedding historical and institutional patterns of bias, particularly against individuals of color.").

[47] *See* Claudia Polsky, *Open Records, Shuttered Labs: Ending Political Harassment of Public University Researchers*, 66 UCLA L. REV. 208, 209 (2019) ("[S]cholars in states with broad open records laws have increasingly received harassing records requests from requesters politically or economically threatened by the intellectual work they seek to reveal. Such requests, impair[] the core intellectual functions of the university. Equally worrisome, harassing record requests chill research on critical contemporary issues--a knowledge-generation role of universities that is essential to a democracy, which depends on an informed citizenry."); *see also* Michael Halpern, *Corporations and Activists Are Exploiting Open Records Laws. California Is Trying to Change That*, UNION OF CONCERNED SCIENTISTS (Feb. 22, 2019, 4:39 PM EST), https://blog.ucsusa.org/michael-halpern/corporations-and-activists-are-exploiting-open records-laws-california-is-trying-to-change-that ("[O]pen records requests have disrupted or derailed the careers of law professors, biologists, tobacco researchers, chemical toxicologists, and many others whose work is found to be inconvenient or objectionable. Some scientists and their families face sustained harassment and even get death.

deep personal profiles.

4. Personal Data Hygiene

A promising way to limit data collection is for individuals to develop better "data hygiene" by taking simple steps to filter the personal information shared with third parties. Five simple suggestions for the average digital device user:

A. Turn off location services in phone settings (and in other devices) when a specific application is not being used. Don't worry, you can always turn the location service back on when needed.

B. Don't allow "contact sharing" between applications. When prompted to share your contact or email list with a new program, follow Nancy Reagan's time-honored advice: "Just say no."

C. Dial back advertising settings in your major social media platforms such as Twitter, Instagram, Gmail and Facebook. All of these platforms have "Privacy Settings" tabs, allowing the user some degree of control over data sharing with third-party advertisers.

D. Limit the use of third-party cookies. You can do this by going into the settings of your web browser and moderate the dropping of cookies on the browser by third parties. It's also good hygiene to wipe all cookies after long periods of time. This may disrupt your log-in of seldom used sites, but that is the tradeoff.

E. Finally, exercise one's privacy rights. The new California Privacy Law, the CCPA, allows you to request a company to tell you the data they have gathered about you, delete some of that day and to opt out of data sharing with third parties. Each company with either $25 million revenue or 50,000 consumer names doing business in the state of California is now obligated to provide a prominent "Opt Out" button on its web site. Any California resident can avail themselves of this window into the data collection practices of myriad firms.

Conclusion

Where Do We Go from Here?

At the outset of this new decade, we find ourselves standing at the intersection of two incoming trains—the explosive growth of Big Data and the rapid development of AI technologies. Will we get caught in this intersection or will we figure out as a society how to harness both trends and use them for the benefit of our culture, economy and planet? In order to achieve the later outcome, our leaders need to thoughtfully define AI without clouding the debate with erroneous fear. Consequently, we need to ask the precise question of how best to implement AI technologies with a view toward enhancing our civil rights and promoting economic progress. At this early stage, it's also appropriate to ask: Who should be framing these questions and making these decisions?

Perhaps this suggestion borders on a truism, but it behooves those interested in addressing this question to think about how we might assemble the best minds and forward-looking thinkers on this topic, drawing not only from the tech world, but from civil society leaders, sociology, economics, politics, law and pure sciences.

AI Ethics Journal

Coping with a "brave new world" of AI is akin to Huxley's dystopian novel.[48] We can either take concerted action to understand and productively apply the new world of AI, or we will find ourselves flattened by the steamroller of technology under the guise of "progress."

Members of the legal profession have a special obligation as custodians of data, because we are viewed by society as "arbiters of truth" and our 21st Century truths increasingly rely on models drawn from data. If the quality of a conclusion is only as good as the quality of the data inputs to be interpreted by an algorithm, then attorneys trained in concepts of evidence and civil procedure must energetically exercise their powers in this role of determining what facts constitute admissible "evidence" for a given case. As any law student taking Evidence might observe, admitting a fact into evidence is not a simple matter as facts must run a complex legal gauntlet before they find themselves before a jury or trier of fact.

As crafters of many of the algorithms that increasingly determine public benefits and detriments, computer scientists and engineers also should feel a heightened sense of responsibility for the outputs of their work. It is inherently difficult for an individual to recognize his or her own "bias" with respect to a matter, even an element of an algorithm that appears to be neutral on its face. Yet the inclusion of certain sets of "neutral" elements can sway a result one way or another. A data point that includes an individual's age, location or education is not so much a single marker as a collection of aggregate facts. Parsing such facts poses a great challenge for the data scientists of the next decade as they train AI programs to incorporate information into their programs. Finally, public policy makers must avoid the temptation to jump to conclusions based on news reports or incomplete studies about the nature or track record of AI programs. As discussed in this article, AI will be imperfect so long as data inputs are flawed or incomplete. AI will yield inaccurate results so long as its programmers incorporate biases, both hidden and overt. Yet the promise of AI to improve human decision making and to crunch data at scales not possible by humans cannot be ignored, as so many of our global problems, ranging from water shortages to climate change to a quest for better energy sources cry out for us to harness all of the tools in our arsenal of thinking, including AI.

At the close of World War II, the advent of a new technology proved that it could both extinguish humanity and perhaps also benefit the world through peaceful harnessing of atomic fission.[49] Seventy-five years later, we are still engaged in debate as to how to balance the destructive power of nuclear arms and nuclear waste with the benefits to society. Artificial Intelligence is not unlike atomic power in that respect, as we are only at the beginning of the long journey toward wisely incorporating AI into our decision-making processes. Just as the incorporation of the computer took decades to integrate into our working systems and personal lives, we should recognize that this perilous road will be marked by both triumphs and mistakes. Utilization of bad data could lead to catastrophic consequences for humans and for our environment. Allowing AI control over vital functions that might be manipulated in ways that promote human suffering and disparity could result in damage to people that cannot be reversed. To that end, readers of this article must reflect on the roles they might play to control, moderate and influence the evolution of this powerful technology, treating it with the awe and gravity that it deserves.

[48] Huxley, Aldous, "Brave New World," Chatto & Windus, 1932.

[49] *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer*, Bird, Kai and Sherwin, Martin, Knopf, 2005.

AI Ethics Journal

References

[1] Brian Barrett, *Lawmakers Can't Ignore Facial Recognition's Bias Anymore*, WIRED (July 26, 2018, 4:59 PM), https://www.wired.com/story/amazon-facial-recognition-congress-bias-law-enforcement/.

[2] Kate Conger et al., *San Francisco Bans Facial Recognition Technology*, N.Y. TIMES (May 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html; *see also* Bruce Schneier, *We're Banning Facial Recognition. We're Missing the Point.*, N.Y. TIMES: THE PRIVACY PROJECT (Jan. 20, 2020), https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html ("Communities across the United States are starting to ban facial recognition technologies. In May [2019], San Francisco banned facial recognition; the neighboring city of Oakland soon followed, as did Somerville and Brookline in Massachusetts (a statewide ban may follow). In December [2019], San Diego suspended a facial recognition program in advance of a new statewide law, which declared it illegal, coming into effect.").

[3] NISTR 8280, Grother, Patrick, Ngan and Hanakoa, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* are intended to inform policymakers and to help software developers better understand the performance of their algorithms. Face recognition technology has inspired public debate in part because of the need to understand the effect of demographics on face recognition algorithms. Full publication available at doi.org/10.6028/NIST.IR.8280.

[4] NIST.gov News: "While it is usually incorrect to make statements across algorithms, we found empirical evidence for the existence of demographic differentials in the majority of the face recognition algorithms we studied," said Patrick Grother, a NIST computer scientist and the report's primary author. "While we do not explore what might cause these differentials, this data will be valuable to policymakers, developers and end users in thinking about the limitations and appropriate use of these algorithms." Reported on NIST.gov, December 19, 2019.

[5] Shepardson, David. Reuters Technology News, November 5, 2019. *In review of fatal Arizona crash, U.S. agency says Uber software had flaws. See also: MIT Media Lab: Algorithmic Justice League Project, available at media.mit.edu.*

[6] NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS, at xiii (2010).

[7] Bernard Marr, *The Key Definitions of Artificial Intelligence (AI) that Explain its Importance,* https://bernardmarr.com/default.asp?contentID=1352

[8] See generally: John Koetsier, *How Amex Uses AI to Automate 8 Billion Risk Decisions (And Achieve 50% Less Fraud),* Forbes (September 21, 2020), https://www.forbes.com/sites/johnkoetsier/2020/09/21/50-less-fraud-how amex-uses-ai-to-automate-8-billion-risk-decisions/#4a1c5b491a97

[9] For an excellent discussion of four types of machines by an AI researcher, see Arend Hintze, *Understanding the Four Types of Artificial Intelligence*, GOV'T TECH. (Nov. 14, 2016), https://www.govtech.com/computing/Understanding-the-Four-Types-of-Artificial-Intelligence.html.

[10] *See* Daisuke Wakabayashi, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam*, N.Y. TIMES: TECH. (Mar. 19, 2018),https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html ("[A]n autonomous car operated by Uber . . . struck and killed a woman on a street in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology [T]he crash in Tempe will draw attention among the general public to self-driving cars, said Michael Bennett, an associate research professor at Arizona State University 'We've imagined an event like this as a huge inflection point for the technology and the companies advocating for it,' he said. 'They have to do a lot to prove that the technology is safe.'").

AI Ethics Journal

[11] *See* Stephen Heyman, *Photos, Photos Everywhere*, N.Y. TIMES (July 29, 2015), https://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html ("The growth in the number of photos taken each year is exponential: It has nearly tripled since 2010 and is projected to grow to 1.3 trillion by 2017. The rapid proliferation of smart phones is mostly to blame."); Amy Hobbs, *[Stats] How Many Photos Have Ever Been Taken?*, FSTOPPERS (Mar. 10, 2012), https://fstoppers.com/other/stats-how-many-photos-have-ever been-taken-5173 (estimating the total number of analogue photographs to be 3.5 trillion).

[12] *See Total Number of Websites*, INTERNET LIVE STATS, https://www.internetlivestats.com/total-number-of-websites/ (last visited Feb. 15, 2020); *How Many Websites Are There Around the World?*, MILL FOR BUS. (Feb. 12, 2020), https://www.millforbusiness.com/how-many-websites-are-there/.

[13] *See* Jennifer Valentino-DeVries et al., *Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret*, N.Y. TIMES (Dec. 10, 2018), https://www.nytimes.com/interactive/2018/12/10/business/location-dataprivacy-apps.html ("At least 75 companies receive anonymous, precise location data from apps whose users enable location services to get local news and weather or other information . . . . Many location companies say that when phone users enable location services, their data is fair game."); *see also* Ryan Gallagher, *NSA Can Reportedly Track Phones Even When They're Turned Off*, SLATE (July 22, 2013, 4:06 PM), https://slate.com/technology/2013/07/nsa can-reportedly-track-cellphones-even-when-they-re-turned-off.html ("[T]o spy on phones when they are turned off, agencies would usually have to infect the handset with a Trojan that would force it to continue emitting a signal if the phone is in standby mode, unless the battery is removed. In most cases, when you turn your phone off . . . it will stop communicating with nearby cell towers and can be traced only to the location it was in when it was powered down.").

[14] *See* Mike Chan, *Big Data in the Cloud: Why Cloud Computing Is the Answer to Your Big Data Initiatives*, THORN TECHS. (Sept. 10, 2018), https://www.thorntech.com/2018/09/big-data-in-the-cloud/ (discussing several key advantages of combining "big data analytics and cloud computing").

[15] *See* Michael Fertik, *Why Your Data Will Never Be Deleted*, Forbes (June 9, 2015, 10:14 AM), https://www.forbes.com/sites/michaelfertik/2015/06/09/why-your-data-will-never-be-deleted/#40fb590a2371.

[16] See generally: John Markoff, *Computer Wins on 'Jeaopardy!': Trivial, It is not.* N.Y. TIMES (February 16, 2011), https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

[17] Jo Best, *IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next,* Tech Republic (September 9, 2013), https://www.techrepublic.com/article/ibm-watson-the-inside story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/.

[18] See generally: IBM and Wellpoint put #ibmwatson to work in Healthcare, IBM, https://www.ibm.com/support/pages/ibm-and-wellpoint-put-ibmwatson-work-healthcare.

[19] *How Watson helps lawyers find answers in legal research: ROSS Intelligence takes Watson to law school*, MEDIUM.COM (Jan. 4, 2017), https://medium.com/cognitivebusiness/how-watson-helps-lawyers-find-answers-in legal-research-672ea028dfb8.

[20] ZDNet, "IBM Watson: What Are Companies Using it For," Conner Forrest, September 1, 2015.

[21] Edge Up Sports, LLC: Using Cognitive technology to help fantasy football "owners" make better roster decisions. IBM, https://www.ibm.com/case-studies/a787535m28346z55.

[22] CNBC, "TikTok Reveals Detailed User Numbers For the First Time," August 24, 2020. This article includes the following data: TikTok revealed specific U.S. and global growth milestones for the first time in a lawsuit against the U.S. government. TikTok has about 100 million monthly active U.S. users, up nearly 800% percent from Jan. 2018. TikTok

AI Ethics Journal

said it has about 50 million daily active U.S. users Here's the breakdown of TikTok's U.S. user growth: January 2018: 11,262,970 U.S. monthly active users (MAUs), February 2019: 26,739,143, October 2019: 39,897,768, June 2020: 91,937,040, August 2020: More than 100 million based on quarterly usage globally, TikTok has experienced similar surges in users. The company said it had about 55 million global users by Jan. 2018. That number ballooned to more than 271 million by Dec. 2018 and 507 million by Dec. 2019. This month, TikTok surpassed 2 billion global downloads and reported nearly 700 million monthly active users in July.

[23] *Kevin Poulsen and* Robert McMillan, *TikTok Tracked User Data Using Tactic Banned by Google,* Wall Street Journal, Online.; Updated Aug. 11, 2020 4:58 pm ET

[24] Wired Magazine, TikTok Finally Explains How the 'For You' Algorithm Works, June 18, 2020.

[25] McNamee, Roger, "Zucked—Waking Up to the Facebook Catastrophe," pp. 90-91.

[26] Fast Company, "How Facebook's 'Like' button Hijacked our Attention and broke the 2010s," by Christopher Zara, December 18, 2019.

[27] Katherine Bindley & Wilson Rothman, *Facebook Has a New Data Policy—Here's the Short Version*, WALL ST. J. (Apr. 20, 2018, 9:29 AM ET), https://www.wsj.com/articles/facebook-has-a-new-data-policyheres-the-short-version 1524230950.

[28] Kashmir Hill, *Turning Off Facebook Location Tracking Doesn't Stop It from Tracking Your Location,* GIZMODO (Dec. 12, 2018, 12:20 PM), https://gizmodo.com/turning-off-facebook-location-tracking-doesnt-stop-it-f-1831149148 (""Facebook does not use WiFi data to determine your location for ads if you have Location Services turned off,' said a Facebook spokesperson by email. 'We do use IP and other information such as check-ins and current city from your profile.'").

[29] Jake Kanter *Facebook Is Tracking You in Ways You Never Knew – Here's the Crazy Amount of Data It Sucks up*, BUS. INSIDER (June 12, 2018, 2:12 AM), https://www.businessinsider.com/facebook-reveals-all-the-way-it-tracks-user behaviour-2018-6.

[30] In particular, Facebook must "expand its privacy protections across Facebook itself, as well as on Instagram and WhatsApp. It must also adopt a corporate system of checks and balances to remain compliant, . . . [and] maintain a data security program, which includes protections of information such as users' phone numbers." Mark Zuckerberg's oversight in privacy and security matters also has been diminished under the FTC settlement. Specifically, Facebook must "create a new privacy committee with independent board members who cannot be removed without a two thirds shareholder vote. Zuckerberg and designated compliance officers each must submit individual quarterly compliance reports to the FTC." Finally, "a third-party assessor will monitor Facebook's privacy-related decisions going forward." Mike Snider & Edward C. Baig, *Facebook Fine $5 Billion by FTC, Must Update and Adopt New Privacy, Security Measures*, USA TODAY (July 24, 2019, 8:54 AM/ET), https://www.usatoday.com/story/tech/news/2019/07/24/facebook-pay-record-5-billion-fine-u-s-privacy violations/1812499001/ ("The $5-billion FTC fine is nearly 20 times greater than the largest privacy or data security penalty that has ever been assessed worldwide and is one of the largest imposed by the U.S. government for any violation. The commission approved the settlement with a 3-2 vote, with the dissenting commissioners wanting tougher action taken against Zuckerberg.")

[31] *Id.* ("[The SEC complaint alleged that] Cambridge Analytica paid an academic researcher to 'collect and transfer data from Facebook to create personality scores for approximately 30 million Americans' and that Facebook discovered this misuse in 2015 but failed to correctly disclose it for more than two years."). [32] The Author's Google Maps Timeline for a one-year period, provided by Alex Alben. For more on this point, see: *Google Is Tracking Your Location—Even Without Your Permission, Report Says,* Fortune Magazine, August 13, 2018; https://fortune.com/2018/08/13/google-tracking-locations-without-permission/

[32] The Author's Google Maps Timeline for a one-year period, provided by Alex Alben. For more on this point, see: *Google Is Tracking Your Location—Even Without Your Permission, Report Says,* Fortune Magazine, August 13, 2018; https://fortune.com/2018/08/13/google-tracking-locations-without-permission/

[33] *See* Janko Roettgers, *Facebook Exec Doesn't Expect Privacy Backlash to Impact Revenue*, VARIETY (Apr. 13, 2018, 9:39 AM PT), https://variety.com/2018/digital/news/facebook-exec-privacy-backlash-revenue-1202752652/ ("[According to Facebook's vice president of global marketing solutions, Carolyn Emerson,] the company hasn't seen many people change their privacy settings on the service. 'People are going in checking it out, for sure,' she said. [But] that curiosity doesn't necessarily result in any changes, at least not for now. 'We are not seeing a surge in any changing in consumer behavior,' she said."); *see also The State of Privacy in Post-Snowden America*, PEW RESEARCH CTR. (Sept. 21, 2016), https://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in america/ [hereinafter *The State of Privacy*] ("Even after news broke about the NSA surveillance programs, few Americans took sophisticated steps to protect their data, and many were unaware of robust actions they could take to hide their online activities.");

[34] Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 320 (2012) ("The mosaic theory requires courts to apply the Fourth Amendment search doctrine to government conduct as a collective whole rather than in isolated steps. Instead of asking if a particular act is a search, the mosaic theory asks whether a series of acts that are not searches in isolation amount to a search when considered as a group. The mosaic theory is therefore premised on aggregation: it considers whether a set of nonsearches [sic] aggregated together amount to a search because their collection and subsequent analysis creates a revealing mosaic.").

[35] Natasha Singer, *Mapping, and Sharing, the Consumer Genome*, N.Y. TIMES (June 16, 2012), https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html ("[Acxiom] is integrating what it knows about our offline, online and even mobile selves, creating in-depth behavior portraits in pixilated detail. Its executives have called this approach a '360-degree view' on consumers.").

[36] *See* VT. OFFICE OF THE ATTORNEY GEN., GUIDANCE ON VERMONT'S ACT 171 OF 2018 DATA BROKER REGULATIONv6 (2018), https://ago.vermont.gov/wp-content/uploads/2018/12/2018-12-11-VT-Data-Broker-Regulation-Guidance.pdf; *see also* Steven Melendez, *A Landmark Vermont Law Nudges over 120 Data Brokers out of the Shadows*, FAST COMPANY (Mar. 2, 2019), https://www.fastcompany.com/90302036/over-120-data-brokers-inch-out-of-the-shadows-under landmark-vermont-law ("The law also requires companies to spell out whether there's any way for consumers to opt out of their data collections, to specify whether they restrict who can buy their data, and to indicate whether they've had any data breaches within the past year.").

[37] Jason Morris & Ed Lavandera, *Why Big Companies Buy, Sell Your Data*, CNN: BUSINESS, https://www.cnn.com/2012/08/23/tech/web/big-data-acxiom/index.html (last updated Aug. 23, 2012, 3:52 PM ET) ("Data is now a $300 billion-a-year industry and employs 3 million people in the United States alone").

[38] For a brief discussion of how the General Data Protection Regulation ("GDPR") effectively prohibits the formation of a personal-data secondary market in Europe, compare Chiara Rustici, *Personal Data and the Next Subprime Crisis*, FORBES (July 24, 2018, 2:17 PM), https://www.forbes.com/sites/chiararustici/2018/07/24/personal-data-and-the-next-subprimecrisis/#6d60080170aa (suggesting that "there will never be a deep and liquid personal data ***secondary*** market" because the GDPR: (1) prohibits re-purposing lawfully-collected personal data;

AI Ethics Journal

(2) absolutely bars utilizing such data "past an agreed time frame"; and (3) heavily restricts "any onward-transfer" of such data— i.e., "personal data may behave as an asset, . . . but it will never behave as a commodity [in Europe]") (emphasis in original).

[39] Adam Schwartz, *Vermont's New Data Privacy Law*, ELECTRONIC FRONTIER FOUND. (Sept. 27, 2018), https://www.eff.org/deeplinks/2018/09/vermonts-new-data-privacy-law ("But it does not address "first-party" data mining . . .. For example, the Vermont law does not cover a social media platform like Facebook, or a retailer like Walmart, when those companies gather information about how consumers interact with their own websites.").

[40] *CCPA and California's New Registration Requirement*, NAT'L L. REV. (Sept. 16, 2019), https://www.natlawreview.com/article/ccpa-and-california-s-new-registration-requirement. The California law defines "data broker" in the same manner as the Vermont law, and it exempts credit reporting agencies (covered by the Fair Credit Reporting Act), financial institutions (covered by the Gramm-Leach-Bliley Act) and covered entities under HIPAA. *Id.*

[41] *See Public Opinion on Privacy*, EPIC.ORG, https://epic.org/privacy/survey/ (citing Sam Sabin, *Most Voters Say Congress Should Make Privacy Legislation a Priority Next Year*, MORNING CONSULT (Dec. 18, 2019, 12:01 AM ET), https://morningconsult.com/2019/12/18/most-voters-say-congress-should-make-privacy-legislation-a-priority next-year/) ("A new poll of registered voters found that 79% of Americans believe that Congress should enact privacy legislation and 65% of voters said data privacy is 'one of the biggest issues our society faces.'").

[42] *See* Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1967, 1972, 1988 (2015) ("Executive agencies frequently release government information under sunshine laws such as the Freedom of Information Act (FOIA), which requires disclosures in response to public records requests provided that no law prohibits the release. Private companies, such as data brokers and app developers, are compiling information from public records, combining it with information from other sources, and repackaging the combined information as new products or services.") (citations omitted); Kirsten Martin & Helen Nissenbaum, *Privacy Interests in Public Records: An Empirical Investigation*, 31 HARV. J.L. & TECH. 111, 120 (2017) ("[C]ommercial stakeholders, such as data brokers, enjoy greater efficiencies in their bulk collection of data from public records, from which they extract knowledge that is attractive to other stakeholders in various sectors.") (citations omitted).

[43] *See, e.g.*, WASH. REV. CODE § 42.56.100 (West, Westlaw through ch. 2 of 2020 Reg. Sess.) ("Agencies shall adopt and enforce reasonable rules and regulations . . . to provide full public access to public records, to protect public records from damage or disorganization . . . ."); *see generally* LOUIS D. BRANDEIS, OTHER PEOPLE'S MONEY: AND HOW THE BANKERS USE IT 91 (Martino Fine Books 2009) (1914) ("Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.").

[44] *See* Nissen v. Pierce Cty., 357 P.3d 45, 55-56 (Wash. 2015) (en banc) (holding that content of work-related text messages sent and received by a county prosecutor on his private cell phone in his official capacity were "public records" under the Public Records Act); West v. Puyallup, 410 P.3d 1197, 1201 (Wash. Ct. App. 2018) (noting that a public official's posts on a personal social media page can constitute an agency's public records subject to disclosure under the Public Records Act if the posts relate to the conduct of government and are prepared within a public official's scope of employment or official capacity).

[45] *See* LEXISNEXIS, TEN COMPELLING REASONS TO RELY ON LEXISNEXIS® PUBLIC RECORDS AS YOU RESEARCH PEOPLE, BUSINESSES AND LOCATIONS 1 (2012), http://www.lexisnexis.com/pdf/Ten%20Reasons_Corp_

AI Ethics Journal

Gov_FINAL.pdf. ("Access more than 36 billion public records—one of the world's largest online collections. [C]ast the broadest net possible to capture the most current and complete picture of a person, business or location. LexisNexis Public Records also offers more types of records, including email, health-care provider sanctions, employment locator[] and others . . . ."); *see also* Ms. Smith, *Cha-Ching of Scraping: Data Brokers Digging up & Selling Your Digital Dirt*, CSO: PRIVACY &SECURITY FANATIC (Oct. 12, 2010, 1:25 PM PST), https://www.csoonline.com/article/2227434/cha-ching-of-scraping--data-brokers digging-up---selling-your-digital-dirt.html ("[S]ite scraping happens all the time, ranging from free do-it-yourself scraping software to screen-scrapers that charge between $1,500 and $10,000 for most jobs. The website PatientsLikeMe.com discovered media-research Nielsen Co. was scraping all messages off the private online forum, messages that were supposed to be viewable only by members who have agreed not to scrape, and not by intruders such as Nielsen.") (citations & internal quotation marks omitted)

[46] *See* Colleen V. Chien & Jason Tashea, *Better Data and Smarter Data Policy for a Smarter Criminal Justice System System*, THE ETHICAL MACHINE (Dec. 10, 2018), https://ai.shorensteincenter.org/ideas/2018/12/10/better data-and-smarter-data-policy-for-a-smarter-criminal-justice-system-system ("Across the country, courts are using profile-based risk assessment tools to make decisions about pretrial detention. The tools are built on aggregated data about past defendants to identify factors that correlate with committing a subsequent crime or missing a trial date. They are used to score individuals and predict if pretrial incarceration is necessary because these tools are built on historical data, they run a real risk of reinforcing the past practices that have led to mass incarceration, like the over incarceration of poor and minority people."); *see also* Colleen V. Chien & Clarence Wardell III, *Make the First Step Act a Smarter Step by Opening the Risk Assessment Black Box*, THE HILL:CRIM.JUST. (Dec. 16, 2018, 8:00 AM EST), https://thehill.com/opinion/criminal-justice/421552-make-the-first-step-act-a-smarter-step-by opening-the-risk?amp ("[T]he reliance on factors like a person's history, educational background, and other demographic factors to classify them risks exacerbating and further embedding historical and institutional patterns of bias, particularly against individuals of color.").

[47] *See* Claudia Polsky, *Open Records, Shuttered Labs: Ending Political Harassment of Public University Researchers*, 66 UCLA L. REV. 208, 209 (2019) ("[S]cholars in states with broad open records laws have increasingly received harassing records requests from requesters politically or economically threatened by the intellectual work they seek to reveal. Such requests, impair[] the core intellectual functions of the university. Equally worrisome, harassing record requests chill research on critical contemporary issues--a knowledge-generation role of universities that is essential to a democracy, which depends on an informed citizenry."); *see also* Michael Halpern, *Corporations and Activists Are Exploiting Open Records Laws. California Is Trying to Change That*, UNION OF CONCERNED SCIENTISTS (Feb. 22, 2019, 4:39 PM EST), https://blog.ucsusa.org/michael-halpern/corporations-and-activists-are-exploiting-open records-laws-california-is-trying-to-change-that ("[O]pen records requests have disrupted or derailed the careers of law professors, biologists, tobacco researchers, chemical toxicologists, and many others whose work is found to be inconvenient or objectionable. Some scientists and their families face sustained harassment and even get death.

[48] Huxley, Aldous, "Brave New World," Chatto & Windus, 1932.

[49] *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer*, Bird, Kai and Sherwin, Martin, Knopf, 2005.