EMORY | LAW

Emory Law Journal

Volume 69 | Issue 4

2020

Defending Data: Toward Ethical Protections and Comprehensive Data Governance

Elizabeth R. Pike

Recommended Citation

Elizabeth R. Pike, *Defending Data: Toward Ethical Protections and Comprehensive Data Governance*, 69 Emory L. J. 687 (2020).
Available at: https://scholarlycommons.law.emory.edu/elj/vol69/iss4/2

DEFENDING DATA: TOWARD ETHICAL PROTECTIONS AND COMPREHENSIVE DATA GOVERNANCE

Elizabeth R. Pike[*]

ABSTRACT

The click of a mouse. The tap of a phone screen. Credit card purchases. Walking through an airport. Driving across a bridge. Our activities, both online and offline, are increasingly monitored, converted into data, and tracked. These troves of data—collected by entities invisible to us, stored in disparate databases—are then aggregated, disseminated, and analyzed. Data analytics—algorithms, machine learning, artificial intelligence—reveal "truths" about us; weaponize our data to influence what we buy or how we vote; and make decisions about everything from the mundane—which restaurant a search engine should recommend—to the significant—who should be hired, offered credit, granted parole.

This Article is the first to chart the terrain of this novel, networked data landscape and articulate the ways that existing laws and ethical frameworks are insufficient to constrain unethical data practices or grant individuals meaningful protections in their data. The United States' sectoral approach to privacy leaves vast swaths of data wholly unprotected. Recently enacted and proposed privacy laws also fall woefully short. And all existing privacy laws exempt de-identified data from coverage altogether—a massive loophole given that the amount of publicly available data sets increasingly render reidentifying de-identified data trivial. Existing ethical frameworks, too, are unable to address the complex ethical challenges raised by networked datasets.

Accordingly, this Article proposes an ethical framework capable of maintaining the public's full faith and confidence in an industry thus far governed by an ethos of "move fast and break things." The CRAFT framework—arising from considerations of the shortcomings of existing ethical frameworks and their inability to meaningfully govern this novel, networked data landscape—establishes principles capable of guiding ethical decision making

[*] J.D., LL.M. Developed as Director of Privacy Policy, U.S. Department of Health and Human Services. I am tremendously grateful to have had the opportunity to workshop this piece at the HHS Data Privacy Symposium, the NIH Department of Bioethics, and with members of the International Association of Privacy Professionals. I am extremely thankful for the insightful contributions of Efthimios Parasidis, Deven McGraw, and Nicholson Price in considering these pressing issues. The findings and conclusions in this article are my own and do not necessarily represent the official position of the U.S. Department of Health and Human Services. Use of official trade names does not mean or imply official support or endorsement.

across the data landscape and can provide a foundation for future legislation and comprehensive data governance.

> *Every day billions of dollars change hands and countless decisions are made on the basis of our likes and dislikes, our friends and families, our relationships and conversations, our wishes and fears, our hopes and dreams. These scraps of data, each one harmless enough on its own, are carefully assembled, synthesized, traded, and sold…. Platforms and algorithms that promise to improve our lives can actually magnify our worst human tendencies…. This crisis is real, it is not imagined, or exaggerated, or crazy. And those of us who believe in technology's power for good must not shrink from this moment.*
>
> Apple CEO, Tim Cook (Oct. 24, 2018)[1]

INTRODUCTION

When Jessie Battaglia started her search for a babysitter, she turned to an online service that held out a tantalizing offer: Advanced artificial intelligence would analyze a candidate's Facebook, Twitter, and Instagram posts and offer an automated risk rating evaluating categories ranging from potential for bullying to having a bad attitude.[2] The black box algorithm returned scores on a five-point scale, but provided no details about why an otherwise promising candidate received a two out of five for bullying.[3] The hiring mother wondered whether the algorithm had perhaps narrowed in on a movie quotation or song lyric, but was nevertheless unnerved. "Social media shows a person's character," Battaglia responded. "So why did she come in at a two and not a one?"[4]

Today, each of us leave behind vast troves of data about ourselves from lives lived online and offline. Every click of our mouse, every website we visit, every social media "like," our movements in public, and even our domestic activities get converted into data. These data points are collected, stored, shared, and aggregated by third parties otherwise invisible to us. They are then subjected to advanced analytic tools. These analytic tools include (1) algorithms, or complex mathematical equations that use data as inputs and produce inscrutable outputs; (2) artificial intelligence, which trains technology to mimic human intelligence; and (3) machine learning and deep learning, which trains machines to independently learn and generate insights from data that powers the next round of insights.

[1] Tim Cook, CEO, Apple, Keynote Address at the 2018 International Conference of Data Protection and Privacy Commissioners (Oct. 24, 2018).

[2] Drew Harwell, *Wanted: The 'Perfect Babysitter.' Must Pass AI Scan for Respect and Attitude*, WASH. POST: THE SWITCH (Nov. 23, 2018, 11:50 AM), https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/.

[3] *Id.*

[4] *Id.*

These advanced data analytic tools sit behind the decisions that govern our lived experience. Data analytics inform decisions that range from the mundane—which restaurant should a search engine recommend—to the more consequential—who should be hired, extended credit, granted government benefits. And today, these analytic tools even inform who should be hired as a babysitter. The problem is that many of these data analytics are unproven, offer results that cannot be verified, and potentially exacerbate existing social inequities.[5] Because of the complexity of these analytic processes, even the data scientists who develop the tools often cannot discern the reasons a particular input gave rise to an output, making outcomes unverifiable. The technologies are also unproven: We do not yet know if a candidate who scores a one on the bullying scale will truly be a better babysitter than a candidate who scores a two. And yet, critical decisions are routinely being made based on metrics like these. Part II of this Article explores this novel, networked data landscape, the potential benefits of networked data, and the concerns raised by large-scale collections of personal data.

The Article argues that to recognize the benefits of networked data, and minimize the potential harms, data decision-makers must be given meaningful guidance. Part III analyzes existing laws, ultimately concluding that existing and proposed laws fail to provide appropriate safeguards. In a world where billions of data decisions are made daily—by those constructing our data landscape and everyone else who exists within it—data decision-makers need a framework capable of supporting ethical decision making that will foster an ethical, and thereby sustainable, data landscape. For reasons discussed in Part IV, existing ethical frameworks are unable to sufficiently guide and, when needed, constrain data decision making in meaningful ways. Accordingly, Part V proposes a new ethical framework—the CRAFT framework—capable of meaningfully guiding ethical data decision making and informing comprehensive data governance.

I. THE NOVEL, NETWORKED DATA LANDSCAPE

As a society, we have stumbled—unwittingly and unaware—into a world where our most intimate activities, both online and offline, are tracked, collected, stored, aggregated, and analyzed. Each data point, innocuous on its

[5] *See* IAPP, BUILDING ETHICS INTO PRIVACY FRAMEWORKS FOR BIG DATA AND AI 2, 4 (2018); *Announcing a Competition for Ethics in Computer Science, with up to $3.5 Million in Prizes*, MOZILLA BLOG (Oct. 10, 2018), https://blog.mozilla.org/blog/2018/10/10/announcing-a-competition-for-ethics-in-computer-science-with-up-to-3-5-million-in-prizes/ ("In recent years, we've watched biased algorithms and broken recommendation engines radicalize users, promote racism, and spread misinformation."); Harwell, *supra* note 1 (noting that AI technologies remain "entirely unproven, largely unexplained and vulnerable to quiet biases").

own, can have unexpected and powerful ramifications when connected with additional data points and subjected to advanced data analytics. These emerging and powerful analytic tools are reshaping the data landscape in ways both powerful and totally unseen. These include:

- algorithms, the complex, computational equations that data are subjected to that produce a definitive, if unverifiable, outcome;[6]
- artificial intelligence, which teaches computational technologies to mimic human intelligence;[7] and
- machine and deep learning, which teaches machines to analyze data, recognize patterns, develop insights, and then learn from those insights.[8]

Today, the sheer volume of data collected about us doubles each year.[9] In ten years, the amount of data we produce will double every twelve hours.[10] We have reached a point where, according to privacy Professor Paul Ohm, "[F]or almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her."[11]

Personal data are big business.[12] Revenues grew from $7.6 billion in 2011 to $35 billion in 2017, with projections of $103 billion by 2027.[13] Location-

[6] *Digital Decisions*, CTR. FOR DEMOCRACY & TECH., https://cdt.org/issue/privacy-data/digital-decisions (last visited Jan. 27, 2020) [hereinafter *Digital Decisions*] ("In its most basic form, an algorithm is a set of step-by-step instructions—a recipe—'that leads its user to a particular answer or output based on the information at hand.'").

[7] NAT. INSTS. OF HEALTH, NIH STRATEGIC PLAN FOR DATA SCIENCE 29 (2018) (defining artificial intelligence as "the power of a machine to copy intelligent human behavior").

[8] *Id.* at 30 (defining machine learning as "a field of computer science that gives computers the ability to learn without being explicitly programmed by humans"); *id.* at 29 (defining deep learning as a "type of machine learning in which each successive layer uses output from the previous layer as input").

[9] Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCIENTIFIC AMERICAN, Feb. 25, 2017, at 3.

[10] *Id.*

[11] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1748 (2010).

[12] *See* Judith Duportail, *I Asked Tinder for My Data. It Sent Me 800 Pages of My Deepest, Darkest Secrets*, GUARDIAN (Sept. 26, 2017), https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold ("Personal data is the fuel of the economy."); 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.

[13] Kari Paul, *What Is Exactis—And How Could It Have Leaked The Data of Nearly Every American?*, MARKET WATCH (June 29, 2018, 4:23 PM), https://www.marketwatch.com/story/what-is-exactisand-how-could-it-have-the-data-of-nearly-every-american-2018-06-28.

targeted advertising alone was valued at $21 billion in 2018.[14] Many of this data is collected, aggregated, and sold by data aggregators who amass large-scale databases on hundreds of millions of people, generally without consumers' knowledge, permission, or input.[15] As of 2013, Equifax, for example, had approximately 75,000 data elements on hundreds of millions of consumers, including whether a consumer purchased a particular soft drink, shampoo, laxative, or yeast infection product, whether they visited the OB/GYN, how many miles they traveled, and the number of whiskey drinks they consumed.[16] Sometimes these data aggregators classify consumers according to financial vulnerability for sale to companies that may be particularly interested in these groups of people, including companies like payday lenders.[17] All of this data aggregation takes place largely obscured from consumer view.[18]

Health data, too, are a marketable asset as data aggregators generate patient dossiers on hundreds of millions of patients.[19] Leading tech companies—Amazon, Google, Apple, and Uber—all have plans to enter the lucrative digital health marketplace.[20] This Part explores the novel, networked data landscape and its implications for our "digital selves: the troves of *data*, the bits of our identities and activities, sprinkled through a myriad of databases."[21] These Sections note that, although these vast troves of data generated by and about us hold out tremendous potential for benefit, they also raise new, and daunting, ethical and societal challenges that subsequent Parts will expand upon in considering the adequacy of existing legal and ethical frameworks.

[14] Valentino-Devries et al., *supra* note 12.

[15] *What Information Do Data Brokers Have on Consumers and How Do They Use It?: Hearing Before the S. Comm. on Commerce, Science, and Transportation*, 113th Cong. 6 (2013) (staff report of Sen. John D Rockefeller, Chairman, S. Comm. on Commerce, Sci., and Transp.).

[16] *Id.* at 15.

[17] *See id.* at 7 ("A number of these products focus on consumers' financial vulnerability, carrying titles such as 'Rural and Barely Making It,' 'Ethnic Second-City Strugglers,' 'Retiring on Empty: Singles,' 'Tough Start: Young Single Parents,' and 'Credit Crunched: City Families.'").

[18] *Id.* at 8.

[19] Adam Tanner, *The Hidden Trade in Our Medical Data: Why We Should Worry*, SCIENTIFIC AMERICAN (Jan. 11, 2017), https://www.scientificamerican.com/article/the-hidden-trade-in-our-medical-data-why-we-should-worry/.

[20] Kirsten Ostherr, *Facebook Knows a Ton About Your Health. Now They Want To Make Money off It.*, WASH. POST (Apr. 18, 2018), https://www.washingtonpost.com/news/posteverything/wp/2018/04/18/facebook-knows-a-ton-about-your-health-now-they-want-to-make-money-off-it/.

[21] Alina Selyukh, *As Amazon Looks To Unlock Your Door, Taking Stock of Meaning of Privacy*, ALL TECH CONSIDERED (Nov. 8, 2017, 9:28 PM), https://www.npr.org/sections/alltechconsidered/2017/11/08/562390160/as-amazon-puts-cameras-in-homes-taking-stock-of-meaning-of-privacy.

A. Our Lives in Data

Our lives—both online and offline—are tracked to a degree many of us do not appreciate. Our keystrokes, our footsteps, our Internet searches, our purchases are all tracked, collected, aggregated, shared, and analyzed.[22] While online, we are tracked by our Internet service providers (ISPs) as we travel across the Internet.[23] ISPs build elaborate—and detailed—profiles of users' browsing histories that they sell to advertisers, political parties, or anyone else without granting users meaningful ways to opt out.[24]

Our online behavior is also tracked by our smartphones. Apps that we download leak our data to third parties.[25] Smartphones with a Verizon Wireless data plan are tracked with a hidden, static, device-specific header that is injected into websites visited, which gives Verizon a complete picture of a customer's Internet browsing, regardless of whether customers use private browsers.[26]

Our browsing habits are tracked across devices—from laptop to desktop to cell phone to tablet—by cookies, small packets of data sent from websites and stored on a user's device.[27] Cookies were developed to track users as they navigated a single website, helping websites "remember" whether users had items in their shopping carts or authenticating users as being permitted to access certain information.[28] Today, tracking has become far more extreme.[29] Users

[22] *Id.* ("We wear step-counting trackers. We document our meals, gatherings and whereabouts online. We let giant tech companies into our homes through voice-activated home assistants.").

[23] Teena Maddox, *The Real Reason Behind the New Law for ISPs and What It Means for Internet Users*, TECH REPUBLIC (Apr. 4, 2017, 8:44 AM), https://www.techrepublic.com/article/the-real-reason-behind-the-new-law-for-isps-and-what-it-means-for-internet-users/.

[24] *Id.*

[25] Andy Greenberg, *An AI that Reads Privacy Policies so that You Don't Have To*, WIRED (Feb. 9, 2018, 7:00 AM), https://www.wired.com/story/polisis-ai-reads-privacy-policies-so-you-dont-have-to/; Jinyan Zang et al., *Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps*, TECH. SCIENCE (Oct. 30, 2015), https://techscience.org/a/2015103001/ (noting that "73% of Android apps shared personal information such as email address with third parties, and 47% of iOS apps shared geo-coordinates and other location data with third parties").

[26] Simson L. Garfinkel & Mary Theofanos, *Non-Breach Privacy Events*, TECH. SCI. (Oct. 9, 2018), https://techscience.org/a/2018100903/.

[27] Ibrahim Altaweel et al., *Web Privacy Census*, TECH. SCI. (Dec. 15, 2015), https://techscience.org/a/2015121502 ("A cookie is a message a web browser (e.g., Internet Explorer, Safari, or Firefox) stores when a website it visits requests it to do so."); Paul, *supra* note 12 ("Today's cookies can link your mobile phone to your laptop, to your home monitoring devices, and much, much more. Creepy? Scary? Orwellian? Yes, yes, yes!").

[28] *See* Paul, *supra* note 12 (defining cookies as "small packets of data sent out by a website when a user visits it and stored in that user's data" that "help the website keep track of the user's movement within the site").

[29] *Id.* ("This tracking has gotten more extreme and detailed in recent years."); Martin Anderson, *72% of 'Anonymous' Browsing History Can Be Attached to the Real User*, STACK (Feb. 7, 2017), https://thestack.com/security/2017/02/07/72-of-anonymous-browsing-history-can-be-attached-to-the-real-user/

who visit the top 100 websites will accumulate over 6,000 cookies.[30] These cookies produce a detailed picture of a user's internet browsing history.[31] As described by Tim Libert of Carnegie Mellon University:

> Companies track you when you visit medical websites, pornography websites, websites for lawyers, websites for politicians, newspaper websites, and the same goes for apps. There are very few things that people don't seek out or share using a computer and nearly all of that is tracked, all the time, by the billion dollar giants you see in the news as well as hundreds of companies you've never heard of.[32]

Social media applications capture far more personal data than is readily apparent. Facebook purchases data from outside parties to learn about users' income and the credit cards they use.[33] Developers of Facebook applications have access to troves of data about the Facebook users who have downloaded their particular applications.[34] Facebook even collects data about those without accounts so that Facebook can target them with advertising as they browse the Internet.[35]

Much of our offline activity—our physical movements through the real world—are similarly tracked and monitored. As we drive, license plate data are captured in real time as cars pass through tolls, travel across bridges, or pass

("Using https connections and VPN services can limit exposure …, though the first method does not mask the base URL of the site being connected to, and the second does not prevent the tracking cookies and other tracking methods which can provide a continuous browsing history."); Jessica Su et al., *De-anonymizing Web Browsing Data with Social Networks* (2017), http://randomwalker.info/publications/browsing-history-deanonymization.pdf ("Web tracking has expanded from simple HTTP cookies to include more persistent tracking techniques, such as the use of flash cookies to 'respawn' or re-instantiate HTTP cookies, the use of cache E-Tags and HTML5 localStorage for the same purpose, and 'cookie syncing' between different third parties. Device fingerprinting attempts to identify users by a combination of the device's properties. New fingerprinting techniques are continually discovered, and are subsequently used for tracking.").

[30] Altaweel et al., *supra* note 27.

[31] Paul, *supra* note 13; Su et al., *supra* note 29.

[32] Nathalie Maréchal, *Targeted Advertising Is Ruining the Internet and Breaking the World*, MOTHERBOARD (Nov. 16, 2018, 1:54 PM), https://www.vice.com/en_us/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world.

[33] Eduardo Porter, *Before Fixing Our Data-Driven Ecosystem, A Crucial Question: How Much Is It Worth?*, N.Y. TIMES, Apr. 18, 2018, at B1.

[34] Robinson Meyer, *My Facebook Was Breached by Cambridge Analytica. Was Yours?*, ATLANTIC (Apr. 10, 2018), https://www.theatlantic.com/technology/archive/2018/04/facebook-cambridge-analytica-victims/557648/ ("Even the developers of rudimentary Facebook apps—like my colleague Ian Bogost, who built a satirical video game on the platform called Cow Clicker— accumulated a massive amount of information about their users, whether or not they intended to. 'If you played Cow Clicker, even just once, I got enough of your personal data that, for years, I could have assembled a reasonably sophisticated profile of your interests and behavior ….'").

[35] Porter, *supra* note 33.

camera-equipped police cars.[36] Our real-world movements are tracked, too, by our smartphones. A blockbuster *New York Times* exposé highlighted the ways our location data are harvested by apps—at intervals of up to 14,000 times a day, accurate to within a few yards—and sold to third parties.[37] Location data can reveal sensitive information, including "whether you've visited a psychiatrist, whether you went to an A.A. meeting, [and] who you might date."[38]

As we travel through the world, we are increasingly tracked using facial recognition technology. In China, facial recognition technology allows people to pay for coffee and withdraw cash from an ATM using their faces.[39] In Europe, facial recognition technology is used at high-end stores and hotels to identify elite customers passing through.[40] In the United States, facial recognition technology is increasingly being used to allow passengers to board airplanes without boarding passes.[41] And police are using facial recognition to cross-check databases in real time to identify those with whom they are speaking.[42]

The things we buy are tracked and analyzed. Data collected about purchases made using grocery store loyalty cards are tracked, analyzed, and sold to third parties.[43] Mastercard and American Express sell data about their customers' purchases to third-party buyers,[44] including to Google, which has access to data covering 70% of all purchases made using credit and debit cards.[45] Google can

[36] Russell Brandom, *Exclusive: ICE Is About to Start Tracking License Plates Across the US*, VERGE (Jan. 26, 2018, 8:04 AM), https://www.theverge.com/2018/1/26/16932350/ice-immigration-customs-license-plate-recognition-contract-vigilant-solutions (describing "a massive vehicle-tracking network generating as many as 100 million sightings per month, each tagged with a date, time, and GPS coordinates of the sighting").

[37] Valentino-Devries et al., *supra* note 11 ("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, The Times found. Several of those businesses claim to track up to 200 million mobile devices in the United States—about half those in use last year. The database … reveals people's travels in startling detail, accurate to within a few yards and in some cases updated more than 14,000 times a day.").

[38] *Id.*

[39] The Week Staff, *How Facial Recognition Technology Is Creeping into Your Life*, WEEK (Nov. 19, 2017), https://theweek.com/articles/737750/how-facial-recognition-technology-creeping-into-life.

[40] *Id.*

[41] *See id.*; Shannon Liao, *Facial Recognition Scans Are Expanding to Delta Flights in Atlanta International Airport*, VERGE (Sept. 20, 2018, 5:55 PM), https://www.theverge.com/2018/9/20/17884476/facial-recognition-scan-delta-flight-atlanta-international-airport.

[42] The Week Staff, *supra* note 39.

[43] James Frew, *How Loyalty Card Apps Compromise Your Privacy*, MAKEUSEOF (May 17, 2017), https://www.makeuseof.com/tag/loyalty-card-apps-compromise-privacy/.

[44] *See* Kate Kaye, *Mastercard, Amex Quietly Feed Data to Advertisers*, ADAGE (Apr. 16, 2013), https://adage.com/article/dataworks/mastercard-amex-feed-data-marketers/240800/.

[45] Michael Reilly, *Google Now Tracks Your Credit Card Purchases and Connects Them to Its Online Profile of You*, MIT TECH. REV. (May 25, 2017), https://www.technologyreview.com/s/607938/google-now-tracks-your-credit-card-purchases-and-connects-them-to-its-online-profile-of-you/.

therefore discern whether advertising served up in internet search results has led consumers to make purchases; Google can use location data, including Google map search and navigation results, to see whether users have traveled to places advertised to them and can then use purchased credit card data to discern what purchases consumers then made.[46]

"Smart" appliances—home voice assistances, smart lights, smart televisions, smart refrigerators, smart thermostats, and smart doorbells[47]—ensure that even behavior that takes place in the confines of our home can be tracked and monitored.[48] The result is entirely new data streams revealing information about our definitionally intimate, domestic activities.[49]

B. Data Analytics Facilitate Important Advances

The data collected about us do not sit unused; they are aggregated and analyzed using powerful tools—algorithms,[50] artificial intelligence,[51] machine learning,[52] and deep learning.[53] Data scientists train machines to independently assess and make connections about complex, networked datasets.[54] These powerful technologies carry with them tremendous possibility and have transformed almost every sector of the economy.[55] Algorithms and machine

[46] *Id.*

[47] Farhad Manjoo, *Your Toaster May Be Watching*, N.Y. TIMES, Oct. 11, 2018, at B1.

[48] *See* Kalev Leetaru, *Even the Data Ethics Initiatives Don't Want to Talk About Data Ethics*, FORBES (Oct. 23, 2018, 3:11 PM), https://www.forbes.com/sites/kalevleetaru/2018/10/23/even-the-data-ethics-initiatives-dont-want-to-talk-about-data-ethics/#2120eeaa1fba ("In the past, the focus was building a product, not collecting data. Televisions focused on giving us the best picture, toasters the best toast and word processors focused on making it as seamless as possible to write prose. Today, building that television involves a conversation around how many ways its cameras, microphones and internet connections can be used to profile its owner, that toaster increasingly phones home about when we eat meals and our culinary tastes, while that word processor builds a literary profile of how and what we write about.").

[49] *See* Kashmir Hill & Surya Mattu, *The House That Spied on Me*, GIZMODO (Feb. 7, 2018, 1:25 PM), https://gizmodo.com/the-house-that-spied-on-me-1822429852 ("[T]he smart home is going to create a new stream of information about our daily lives that will be used to further profile and target us…. Our homes could become like internet browsers, with unique digital fingerprints, that will be mined for profit just like our daily Web surfing is. If you have a smart home, it's open house on your data.").

[50] *Digital Decisions*, *supra* note 6 ("In its most basic form, an algorithm is a set of step-by-step instructions—a recipe—'that leads its user to a particular answer or output based on the information at hand.'").

[51] NAT. INSTS. OF HEALTH, *supra* note 7, at 29 (defining artificial intelligence as "the power of a machine to copy intelligent human behavior").

[52] *Id.* at 30 (defining machine learning as "a field of computer science that gives computers the ability to learn without being explicitly programmed by humans").

[53] *Id.* at 29 (defining deep learning as a "type of machine learning in which each successive layer uses output from the previous layer as input").

[54] *Digital Decisions*, *supra* note 5 ("Computers are able to process very complex algorithms and very large inputs in microseconds, producing what can be opaque and often significant algorithmic decisions.").

[55] *Id.*

learning have been used to improve weather forecasts, fine-tune internet search results, and detect credit card fraud.[56]

Big data analytics have also led to remarkable advances in health care. Researchers who previously had to conduct costly randomized-controlled trials can now query existing datasets to generate insights about our lives, behavior, and health.[57] Machines can identify metastatic breast cancer with 100% accuracy—consistently outperforming trained pathologists.[58] Machines can identify with greater than 96% accuracy whether a six-month-old infant will develop autism at twenty-four months, holding out the promise of earlier intervention.[59] Algorithms and machine learning will increasingly be used in medical care to optimize use of scarce resources.[60] In the near-term, big data

[56] *Id.* ("Almost every sector of the economy has been transformed in some way by algorithms. Some of these changes are upgrades, benefiting society by predicting factual outcomes more accurately and efficiently, such as improved weather forecasts. Other algorithms empower tools, such as Internet search engines, that are indispensable in the information age. These advancements are not limited to traditionally computer-powered fields. Algorithms can help doctors read and prioritize X-rays, and they are better and faster than humans at detecting credit card fraud. Wall Street fortunes depend on who can write the best trade-executing algorithm.").

[57] NAT. INSTS. OF HEALTH, *supra* note 6, at 2 ("Advances in storage, communications, and processing have led to new research methods and tools that were simply not possible just a decade ago."); Adam Rogers & Megan Molteni, *Google's Health Spinoff Verily Joins the Fight Against PTSD*, WIRED (Aug. 7, 2017, 7:00 AM), https://www.wired.com/story/google-verily-aurora-ptsd/ (describing a study in which eligible participants get a "wearable that captures data like heart rate, skin electrical conductivity, and movement" and "an experimental app on their smartphones" to "pick up early signs and symptoms of psychiatric disorders"); Sam Volchenboum, *Social Networks May One Day Diagnose Disease—But at a Cost*, WIRED (June 26, 2017, 10:04 AM), https://www.wired.com/story/social-networks-may-one-day-diagnose-disease-but-at-a-cost/ ("The world is becoming one big clinical trial. Humanity is generating streams of data from different sources every second. And this information, continuously flowing from social media, mobile GPS and wifi locations, search history, drugstore rewards cards, wearable devices, and much more, can provide insights into a person's health and well-being.").

[58] Jennifer Bresnick, *MIT Uses Deep Learning to Create ICU, EHR Predictive Analytics*, HEALTH IT ANALYTICS (Aug. 22, 2017), https://healthitanalytics.com/news/mitusesdeeplearningtocreateicuehrpredictiveanalytics (citing Jennifer Bresnick, *Deep Learning Network 100% Accurate at Identifying Breast Cancer*, HEALTH IT ANALYTICS (May 12, 2017), https://healthitanalytics.com/news/deep-learning-network-100-accurate-at-identifying-breast-cancer).

[59] Robert W. Emerson et al., *Functional Neuroimaging of High-Risk 6-Month-Old Infants Predicts a Diagnosis of Autism at 24 Months*, 9 SCI. TRANSLATIONAL MED. 2882, 2882 (2017).

[60] *See* Ruben Amarasingham et al., *Implementing Electronic Health Care Predictive Analytics: Considerations and Challenges*, 33 HEALTH AFF. 1148, 1148 (2014); David W. Bates et al., *Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients*, 33 HEALTH AFF. 1123, 1124 (2014); Danton S. Char et al., *Implementing Machine Learning in Health Care—Addressing Ethical Challenges*, 378 NEW ENG. J. MED. 981, 981 (2018) ("Private companies are rushing to build machine learning into medical decision making, pursuing both tools that support physicians and algorithms designed to function independently of them."); I. Glenn Cohen et al., *The Legal and Ethical Concerns That Arise From Using Complex Predictive Analytics in Health Care*, 33 HEALTH AFF. 1139, 1139 (2014); Jennifer Bresnick, *MIT Uses Deep Learning to Create ICU, EHR Predictive Analytics*, HEALTH IT ANALYTICS (Aug. 22, 2017), https://healthitanalytics.com/news/mitusesdeeplearningtocreateicuehrpredictiveanalytics (noting that researchers believe "deep learning can underpin a new generation of predictive analytics and clinical decision support tools that will safeguard patients

analytics will increasingly enable us to identify disease earlier and treat disease better.

Data analytics are also deployed to address pressing public health challenges. Text mining of social media posts was able to detect polio and Ebola outbreaks sooner than traditional public health surveillance methodology.[61] Data scientists are training machines to analyze risk factors associated with the opioid epidemic to identify those at risk of becoming opioid dependent.[62] Analytic tools applied to data shared population wide can help identify outbreaks sooner, deploy resources more effectively, and make important strides in advancing the public's health.

Machines may soon play a prominent role in diagnosing disease by analyzing social media posts.[63] Machines can identify markers of depression in the filters selected for Instagram posts, outperforming practitioners' average diagnostic success rate.[64] Facebook and other social media platforms are using text-mining artificial intelligence to identify users at risk of self-harm.[65] Machines have been able to "predict with 80 to 90 percent accuracy whether or not someone will attempt suicide, as far off as two years in the future."[66] It will not be long before our smart phones—detecting skipped trips to the gym, decreasing step counts, or ignored calls and texts from friends—will diagnose mental states, holding out the possibility of increased individual wellbeing.[67]

in the intensive care unit and improve how EHRs function for decision-making"). Algorithms can even be used to predict death. Ravi B. Parikh, *Can a Machine Predict Your Death?*, SLATE (Mar. 13, 2017, 7:15 AM), https://slate.com/technology/2017/03/machines-are-getting-better-at-predicting-when-patients-will-die.html.

[61] *See* Aranka Anema et al., *Digital Surveillance for Enhanced Detection and Response to Outbreaks*, 14 LANCET INFECTIOUS DISEASE 1035, 1036 (2014).

[62] Sanket Shah, *How Predictive Analytics Can Help Address the Opioid Crisis*, HEALTHCARE INFORMATICS (2018), https://www.hcinnovationgroup.com/population-health-management/article/13029027/how-predictive-analytics-can-help-address-the-opioid-crisis; Dennis Wei & Fredrik D. Johansson, *Fighting the Opioid Epidemic with Interpretable Causal Estimation of Individual Treatment Effect*, MEDIUM (Oct. 9, 2018), https://medium.com/@MITIBMLab/fighting-the-opioid-epidemic-with-interpretable-causal-estimation-of-individual-treatment-effect-2b2e68ce69d5.

[63] Volchenboum, *supra* note 56 ("It's now entirely conceivable that Facebook or Google—two of the biggest data platforms and predictive engines of our behavior—could tell someone they might have cancer before they even suspect it.").

[64] Andrew G. Reece & Christopher M. Danforth, *Instagram Photos Reveal Predictive Markers of Depression*, 6 EPJ DATA SCI. 15 (2017).

[65] Megan Molteni, *Artificial Intelligence Is Learning to Predict and Prevent Suicide*, WIRED (Mar. 17, 2017, 7:00 AM), https://www.wired.com/2017/03/artificial-intelligence-learning-predict-prevent-suicide/ ("Facebook will make the option to report the post for 'suicide or self injury' more prominent on the display. In a personal post, Mark Zuckerberg described how the company is integrating the pilot with other suicide prevention measures, like the ability to reach out to someone during a live video stream.").

[66] *Id.*

[67] *Id.*

C. Data Analytics Raise Serious Concerns

Although the data analytics that increasingly govern our world hold out tremendous potential for benefit, the massive scale of networked data collections also raises serious concerns. The sheer number of data points, the ways in which disparate data points are connected, the ease with which data are transmitted, and the inferences that can be generated make any decision to collect a given data point fraught.[68] The Sections that follow analyze the complexities of this novel, networked data landscape and the limited options individuals have to meaningfully opt out.

1. Data are Managed by Unaccountable Third Parties

One of the key features of this novel, networked data landscape is that so much of the data surrendered by us, unknowingly and unwittingly, is collected by entities with whom we have no obvious or direct relationship and used in ways that are hidden from us. Data in this novel, networked data landscape are collected by unaccountable third parties whose interests are not necessarily aligned with ours. As described by Nathalie Maréchal, the Silicon Valley data collection business model is:

> [F]irst, grow the user base as quickly as possible without worrying about revenue; second, collect as much data as possible about the users; third, monetize that information by performing big data analytics in order to show users advertising that is narrowly tailored to their demographics and revealed interests; fourth, profit.[69]

The result is that companies with which you interact have an interest in collecting as much data about you as possible and distributing that information as widely as possible for as much money as possible.[70]

[68] *See* Jacob Brogan, *FTC Report Details How Big Data Can Discriminate Against the Poor*, SLATE (Jan. 7, 2016, 2:20 PM), https://slate.com/technology/2016/01/ftc-report-shows-big-data-can-discriminate-against-the-poor.html; Michael Kassner, *5 Ethics Principles Big Data Analysts Must Follow*, TECH REPUBLIC (Jan. 2, 2017, 6:00 AM), https://www.techrepublic.com/article/5-ethics-principles-big-data-analysts-must-follow/ ("At this point in our history … we can process exabytes of data at lightning speed, which also means we have the potential to make bad decisions far more quickly, efficiently, and with far greater impact than we did in the past."); Michael Zimmer, *OKCupid Study Reveals the Perils of Big-Data Science*, WIRED (May 14, 2016, 7:00 AM), https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/ ("The most important, and often least understood, concern is that even if someone knowingly shares a single piece of information, big data analysis can publicize and amplify it in a way the person never intended or agreed.").

[69] Leetaru, *supra* note 48 ("As raw data itself becomes the lifeblood of the modern digital world, more and more companies are built not around providing a neutral service like a word processor, but rather around the collection, exploitation and monetization of data, with services becoming merely portals through which to acquire and act upon such data."); Maréchal, *supra* note 31.

[70] Maréchal, *supra* note 32, at 5 ("Just like 20th century firms like General Motors and Ford invented

The sheer scale of this data-collecting endeavor was made clear through the 2018 discovery of an unsecured data set collected by the marketing firm Exactis,[71] a shadowy company with only ten known employees.[72] By tracking people using internet cookies, Exactis amassed data on over 230 million consumers and 110 million businesses.[73] Exactis had entries that included more than 400 variables on individual subjects, including "whether the person smokes, their religion, whether they have dogs or cats, and interests as varied as scuba diving and plus-size apparel."[74] A hidden third party, with whom few users have ever knowingly transacted, nevertheless knew facts in nearly 400 categories about almost every American, and could choose to make that information available to the highest bidder—whenever and however they saw fit.[75]

Third parties amass data by combining disparate datasets to generate a more complete picture of individuals, regardless of whether any individual wants that information shared or analyzed. Facebook, for example, launched a project to combine anonymized patient records collected from hospitals with Facebook profiles to create digital health profiles of Facebook users.[76] Although people consent to share medical information with a hospital, and may choose to share information about their lives on Facebook, they might nevertheless prefer that the social media giant does not also know their intimate health details. And yet, they were given no opportunity to opt out.

Third-party data collectors do not necessarily share the same sensitivities about data. For example, dating app Grindr shared information about its users' HIV status, along with email address and GPS location, with outside companies that help Grindr optimize the app's functionality.[77] Despite the sensitivity of this

mass production and managerial capitalism, Google and Facebook figured out how to commodify 'reality' itself by tracking what people (and not just their users) do online (and increasingly offline too), making predictions about what they might do in the future, devising ways to influence behavior from shopping to voting, and selling that power to whoever is willing to pay.").

[71] Andy Greenberg, *Marketing Firm Exactis Leaked a Personal Info Database with 340 Million Records*, PRIVACY BLOG (June 27, 2018, 1:34 PM), https://privacyblog.com/2018/06/27/marketing-firm-exactis-leaked-a-personal-info-database-with-340-million-records/.

[72] Paul, *supra* note 12.

[73] Greenberg, *supra* note 71; Paul, *supra* note 12.

[74] Greenberg, *supra* note 71.

[75] *Id.* ("Each record contains entries that go far beyond contact information and public records to include more than 400 variables on a vast range of specific characteristics.").

[76] Ostherr, *supra* note 20.

[77] Azeen Ghorayshi & Sri Ray, *Grindr Is Letting Other Companies See User HIV Status and Location Data*, BUZZFEED NEWS (Apr. 2, 2018, 11:13 PM), https://www.buzzfeednews.com/article/azeenghorayshi/grindr-hiv-status-privacy.

information, the information was shared as plain text, without encryption.[78] Researchers in Denmark publicly released data from nearly 70,000 OkCupid users—including user name, location, and type of relationship (or sex) they were interested in—reasoning that because users agreed to share this information with the company and with other potential daters they did not express a privacy interest in this information.[79] Companies chose to share this data for pragmatic reasons—to increase app functionality or generate research opportunities. Users whose HIV status or sexual preferences were shared could nevertheless feel betrayed that their choices to share data with a presumed finite number of people resulted in making these relatively private facts more public than they ever intended.

2. Large-Scale, Networked Databases Are Vulnerable to Attack

Amassing large-scale, networked databases raises the stakes of any data breach.[80] In September 2017, a data breach of the credit monitoring company Equifax exposed the names, Social Security numbers, and dates of birth of 143 million individuals—nearly half of all Americans.[81] Being included in large-scale databases is a fact of modern life; engaging in financial transactions means being included in databases like Equifax's.

Because individuals have no meaningful way of opting out, every person's vulnerability to identity theft is in the hands of these database developers. And companies that hold large-scale datasets do not always take that responsibility as seriously as they should. In fact, the cybersecurity posture that gave rise to the Equifax breach—the failure to implement a security patch on open-source software—is still in effect in comparable databases; for that reason, the next large-scale data breach has likely already occurred.[82]

The size of the Equifax database made it an attractive target to those who wished to do ill; in fact, because the Equifax data have not yet shown up for sale

[78] *Id.*

[79] Zimmer, *supra* note 68.

[80] Metcalf et al., *Perspectives on Big Data, Ethics, and Society*, COUNCIL FOR BIG DATA, ETHICS, & SOC'Y (May 23, 2016), https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/ ("[T]he emergent properties of massive, connected, and heterogeneous datasets are different than those of "traditional" datasets that remain restricted to a context much closer to their original point of collection.").

[81] Taylor Armerding, *Equifax Breach: Catastrophic, But No Game Changer Yet,* FORBES (Sept. 11, 2018, 12:13 PM), https://www.forbes.com/sites/taylorarmerding/2018/09/11/equifax-breach-catastrophic-but-no-game-changer-yet/; Sean Gallagher, *Equifax Breach Exposed Millions of Driver's Licenses, Phone Numbers, Emails*, ARS TECHNICA (May 8, 2018, 11:13 AM), https://arstechnica.com/information-technology/2018/05/equifax-breach-exposed-millions-of-drivers-licenses-phone-numbers-emails/.

[82] Armerding, *supra* note 81.

on the dark web, analysts suspect the data exfiltration was conducted by a nation state's spy operation.[83]

In March 2018, the Facebook-Cambridge Analytica breach highlighted nonfinancial vulnerabilities in large-scale, networked databases. Facebook previously granted app developers access to user data and permission to request access to users' friends' data as well.[84] In 2013, an app called "thisisyourdigitallife" harvested personal data from almost 300 thousand users and millions of their friends.[85] As a result, data from 87 million Facebook profiles were harvested by Cambridge Analytica and used to form "psychographic" profiles of voters that informed advertising purchases about Brexit, Senator Ted Cruz's presidential primary run, and President Trump's 2016 presidential campaign.[86]

The Facebook-Cambridge Analytica breach made clear that large-scale data breaches can be about more than financial fraud or identity theft. Data collected can be weaponized to target individuals and influence behavior, with resounding geopolitical implications. Moreover, with large-scale databases, assessing the scope of the data breach can be time-consuming and difficult. Facebook initially estimated that 50 million people had been impacted, but subsequently raised the estimate to 87 million.[87] The size and complexity of networked databases render the challenges of forensic cybersecurity analyses a feature, not a bug. Finally, once data have been extracted, there is little that Facebook or any affected user can do—once personal data are outside Facebook's control the data cannot easily be retracted.[88] With massive, networked databases, therefore, potential consequences are far-reaching and not easily managed.

[83] Ryan Whitwam, *The Equifax Breach Might Have Been a Foreign Intelligence Operation*, EXTREME TECH (Feb. 15, 2019, 10:02 AM), https://www.extremetech.com/internet/285827-the-equifax-breach-might-have-been-a-foreign-intelligence-operation.

[84] Sam Meredith, *Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal*, CNBC (Apr. 10, 2018, 9:51 AM), https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html.

[85] *Id*.; Alvin Chang, *The Facebook and Cambridge Analytica Scandal, Explained with a Simple Diagram*, VOX (May 2, 2018, 3:35 PM), https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

[86] Alex Hern & David Pegg, *Facebook Fined For Data Breaches in Cambridge Analytica Scandal*, GUARDIAN (Jul. 10, 2018, 7:01 PM), https://www.theguardian.com/technology/2018/jul/11/facebook-fined-for-data-breaches-in-cambridge-analytica-scandal; Meyer, *supra* note 33.

[87] Meredith, *supra* note 84 ("In an explosive expose [sic] published in mid-March, The Guardian and The New York Times initially reported that 50 million Facebook profiles were harvested for Cambridge Analytica in a major data scandal. This number was later revised to as many as 87 million Facebook profiles."); Meyer, *supra* note 33.

[88] Meyer, *supra* note 34.

3. Data Analytics Reveal Unexpected "Truths"

Big data analytics make predictions about us with consequences ranging from the mild and humorous to the serious and life-threatening.[89] By collecting and analyzing disparate data points, machines may identify traits about individuals that individuals themselves had not known or wanted shared.[90] These findings may be innocuous, like a finding that those who "liked" curly fries on Facebook tended to be smarter.[91] Analyzing a user's Facebook "likes," however, can also reveal sexual orientation, race, political affiliation, immigration status, or other attributes that people could prefer to keep private.[92]

Researchers at Stanford University developed an algorithm that uses facial recognition technology to guess someone's sexual orientation with a greater degree of accuracy than can humans.[93] In countries where homosexuality is illegal, these analyses could place people at serious risk.[94]

Data aggregation can reveal "truths" that were hidden in plain sight. In 2018, fitness-tracking app Strava released "heat maps" of its users' locations.[95] Aggregating the exercise routes of all its users inadvertently revealed sensitive military information, including apparent locations of secret U.S. military bases in Russia, Afghanistan, and Turkey.[96] The level of detail was sufficient to reveal even the internal layouts of several military bases.[97] This data release made clear that "data collection that may have seemed harmless in isolation could upend

[89] W. Nicholson Price II & I. Glenn Cohen, *Privacy in the Age of Medical Big Data*. 25 NATURE MED. 37, 37 (2019); *Digital Decisions*, *supra* note 5 ("Algorithms play a central role in modern life, determining everything from search engine results and social media content to job and insurance eligibility. Unprecedented amounts of information fuel engines that help us make choices about even mundane things, like what restaurant to visit.").

[90] Metcalf et al., *supra* note 80 ("[B]ig data's central power and peril is the ability to network and re-analyze datasets from highly disparate contexts—often in concert—to generate unanticipated insights.").

[91] Brogan, *supra* note 68.

[92] Garfinkel & Theofanos, *supra* note 26.

[93] Heather Murphy, *Why Stanford Researchers Tried to Create a 'Gaydar' Machine*, N.Y. TIMES (Oct. 9, 2017), https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html ("So to call attention to the privacy risks, [Dr. Kosinski] decided to show that it was possible to use facial recognition analysis to detect something intimate, something 'people should have full rights to keep private.' After considering atheism, he settled on sexual orientation.").

[94] The Week Staff, *supra* note 39.

[95] Aja Romero, *How a Fitness App Revealed Military Secret—and the New Reality of Data Collection*, VOX (Feb. 1, 2018, 11:30 AM), https://www.vox.com/technology/2018/2/1/16945120/strava-data-tracking-privacy-military-bases.

[96] *Id.* ("As it turns out, when you put enough soldiers in one place and they exercise in the same locations every day, their collective composite heat map can reveal things over time that no one was expecting.").

[97] *Id.*

closely guarded government secrets" and served as "a wakeup call" to those who had not previously considered the consequences of data analytics.[98]

Data collected by and about us have the power to shape the way the world gets reflected back to us. Leading technology companies, including Google and Facebook, use data analytics to assess a user's personal preferences and tailor advertising and search results accordingly.[99] Research results are personalized, returning results that will generate the most engagement or advertising revenue rather than any objective truth.[100] Using our data, advertisers sell ads to us that entice us to modify our behavior in ways that are precisely targeted to our vulnerabilities.[101]

4. *Data Analytics Can Exacerbate Inequities*

Data analytics, through their seeming ability to distill complex data sets into inscrutable outcomes, are increasingly being used to make decisions about individual lives. Decisions that are increasingly being automated range from the inconsequential—which restaurant a search engine should recommend—to the serious, including who should be hired, extended lines of credit, or granted government benefits.[102] Data analytics can therefore have serious consequences for individuals directly impacted and for society writ large by generating outcomes "without providing an explanation or an opportunity to challenge the decision or the reasoning behind it."[103]

Algorithms—complex mathematical equations—reduce large quantities of data to a singular output.[104] Although algorithmic results appear objective and

[98] *Id.*

[99] Melissa Hammer, *No More Secrets: Gmail and Facebook Can Determine Your Political Values*, TECH. SCI. (Sept. 1, 2015), https://techscience.org/a/2015090105/.

[100] *See* Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCI. AM. (Feb. 25, 2017), https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ ("[I]n the end, all you might get is your own opinions reflected back at you."); Maréchal, *supra* note 31 ("Targeted advertising causes us to experience the internet, and therefore the world, in different ways based on what the surveillance capitalism assemblage thinks it knows about us. This not a recipe for fairness, equality, or a just society.").

[101] Maréchal, *supra* note 32 ("Google and Facebook figured out how to commodify 'reality' itself by tracking what people (and not just their users) do online (and increasingly offline too), making predictions about what they might do in the future, devising ways to influence behavior from shopping to voting, and selling that power to whoever is willing to pay.").

[102] *Digital Decisions*, *supra* note 6.

[103] *Id.*

[104] *Id.* ("Algorithms are essentially mathematical equations. However, unlike mathematical equations you may be familiar with from primary school, algorithmic outputs do not necessarily represent a 'right answer,' defined by an objective truth. Imperfect data sets and human value judgements [sic] shape automated decisions

neutral, algorithms train on datasets that incorporate existing human biases and reflect the hidden biases of their creators.[105] Algorithmic decision-making may therefore reproduce, exacerbate, or amplify biases that already exist in society,[106] negatively impacting groups of people already subject to discrimination.[107]

In some instances, the disparate impacts of algorithmic decision-making are mild. For example, Boston released a mobile application that allowed residents to report potholes directly using their phone's GPS coordinates.[108] Routes traveled by those more likely to own smartphones were more often reported and repaired, resulting in affluent neighborhoods receiving disproportionate pothole repairs and lower-income neighborhoods having potholes under-repaired.[109]

Networking site LinkedIn offered "corrections" for female names—for example, suggesting "Stephen Williams" for "Stephanie Williams"—but did not do so when the genders were reversed.[110] LinkedIn's algorithm trained on word frequency without considering that American men are more likely to have a common name than American women.[111] But this disparity in a social network centered around career advancement and connections could inadvertently steer opportunities away from women.

in intentional and unintentional ways.").

[105] Kate Crawford, *AI's White Guy Problem*, N.Y. TIMES, June 26, 2016, at SR11 ("Like all technologies before it, artificial intelligence will reflect the values of its creators. So inclusivity matters—from who designs it to who sits on the company boards and which ethical perspectives are included. Otherwise, we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes."); Brogan, *supra* note 67; *Digital Decisions*, *supra* note 5 ("[A]lgorithms are imbued with the values of those who create them."); A.R. Lange & Natasha Duarte, *Understanding Bias in Algorithmic Design*, DEM+ND (Apr. 15, 2017), https://demandasme.org/understanding-bias-in-algorithmic-design/ ("Behind every data-driven decision lies a series of human judgments. Decisions about what variables to use, how to define categories or thresholds for sorting information, and which datasets to use to build the algorithm can all introduce bias.").

[106] Char et al., *supra* note 60, at 981–82 ("Algorithms introduced in nonmedical fields have already been shown to make problematic decisions that reflect biases inherent in the data used to train them."); Brogan, *supra* note 67 ("At its worst, big data can reinforce—and perhaps even amplify—existing disparities, partly because predictive technologies tend to recycle existing patterns instead of creating new openings.").

[107] Crawford, *supra* note 104 ("Sexism, racism and other forms of discrimination are being built into the machine-learning algorithms that underlie the technology behind many 'intelligent' systems that shape how we are categorized and advertised to."); *Digital Decisions*, *supra* note 5 (recognizing that "automated decision-making systems can have disproportionately negative impacts on minority groups by encoding and perpetuating societal biases").

[108] *Digital Decisions*, *supra* note 6.

[109] *Id.*

[110] Lange & Duarte, *supra* note 105.

[111] *Id.*

More worrying, as privacy guru Dr. Latanya Sweeney discovered, online searches for a person's name were more likely to show advertisements suggesting that the searched-for person had an arrest record when the name searched was more associated with African Americans than with whites, regardless of whether the searched-for person had actually been arrested.[112] The opaque nature of these algorithms means that no satisfying justification for this disparity has been provided.[113]

Other uses of algorithms result in serious lived consequences.[114] Algorithms are deployed in decision-making that affects critical facets of our lives, including access to health care, credit, insurance, employment, and government programs.[115] As we browse the internet, websites draw conclusions about visitors and assign "e-credit" scores that, based on browsing history and "friends" in social networks, determine whether users will be shown advertising for high-interest or low-interest cards.[116] Dr. Latanya Sweeney has found a racial disparity in the credit cards advertised, with users of color potentially being shown disadvantageous credit cards.[117]

Algorithms are increasingly being used in hiring decisions.[118] In 2014, Amazon decided to automate its hiring process by developing algorithms trained using data about past job applicants.[119] Because of existing gender disparities among its previous job applicants, the algorithm quickly learned to downgrade anything with the word "women's" as a descriptor and upgrade "macho verbs" like "executed" and "captured."[120] Skills that were actually being sought, like

[112] LATANYA SWEENEY, DISCRIMINATION IN ONLINE AD DELIVERY 4 (2013).

[113] *See id.* at 34.

[114] *Digital Decisions*, *supra* note 6 ("Some of the most crucial determinations affecting our livelihoods—such as whether a person is qualified for a job, is creditworthy, or is eligible for government benefits—are now partly or fully automated. In the worst case scenario, automated systems can deny eligibility without providing an explanation or an opportunity to challenge the decision or the reasoning behind it. This opacity can leave people feeling helpless and discourage them from participating in critical institutions.").

[115] *Id.*; Brogan, *supra* note 68.

[116] *Digital Decisions*, *supra* note 6 ("Users who visit a credit card website don't know they're being scored or the criteria or formula behind the score, yet these scores determine their credit opportunities."); *see* Sam Biddle, *Thanks to Facebook, Your Cellphone Company Is Watching You More Closely than Ever*, INTERCEPT (May 20, 2019, 12:50 PM), https://theintercept.com/2019/05/20/facebook-data-phone-carriers-ads-credit-score/.

[117] *See generally* Latanya Sweeney, *Online Ads Roll the Dice*, FED. TRADE COMMISSION (Sept. 25, 2014, 3:59 PM), https://www.ftc.gov/news-events/blogs/techftc/2014/09/online-ads-roll-dice.

[118] Brogan, *supra* note 68; *Digital Decisions*, *supra* note 6.

[119] Jordan Weissmann, *Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.*, SLATE (Oct. 10, 2018, 4:52 PM), https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.

[120] *Id.*

coding, were considered neutral because they appeared in every resume.[121] In 2017, Amazon lost hope in the algorithm's ability *not* to discriminate, and ultimately "shuttered the effort."[122] A critical takeaway, however, is that a leading tech company could not automate hiring in ways that would not increase discrimination. At a time when companies around the world are increasingly looking to automate hiring decisions, it seems unlikely that less tech-savvy companies would be able to do so without discriminating when Amazon could not.[123]

Perhaps most troublingly, algorithms are being used in the criminal justice system to predict criminal recidivism, with algorithms that have demonstrable, disparate racial impacts.[124] Algorithms are used at all stages of the criminal justice system to perform risk assessments about the likelihood that a criminal defendant will re-offend.[125] The allure of an objective risk assessment is apparent: Removing human biases from the assessment of who is likely to reoffend holds out the promise of a more fair, just, and equitable criminal justice system.[126] And yet, a ProPublica investigation discovered that the algorithm produced biased results. The score was unreliable in predicting who would commit violent crime and was twice as likely to mistakenly deem black defendants as high risk of future recidivism, and twice as likely to incorrectly deem white defendants low risk.[127]

Algorithms are increasingly being deployed as cities strive to make their police forces "smart." Cities around the United States—New York, Los Angeles, Chicago, and Miami—are deploying "predictive policing" algorithms to identify

[121] *Id.*

[122] *Id.*

[123] *Id.*("[A]t a time when lots of companies are embracing artificial intelligence for things like hiring, what happened at Amazon really highlights that using such technology without unintended consequences is hard. And if a company like Amazon can't pull it off without problems, it's difficult to imagine that less sophisticated companies can.").

[124] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; *Digital Decisions*, *supra* note 6.

[125] Angwin et al., *supra* note 124.

[126] *Id.*

[127] *Id*. ("In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants."); Crawford, *supra* note 107 (noting that "widely used software that assessed the risk of recidivism in criminals was twice as likely to mistakenly flag black defendants as being at a higher risk of committing future crimes. It was also twice as likely to incorrectly flag white defendants as low risk").

"crime hot spots," sending additional police to these locations.[128] The risk, however, is that algorithms learn from datasets that have socioeconomic biases incorporated.[129] Policing algorithms are trained on data sets that incorporate disparate arrest rates from areas that are already over-policed; the algorithms thus learn to deploy additional officers to areas deemed to be high crime, as inferred from the higher arrest rates, which further increases police presence, increases the number of arrests, and continues to train the algorithm to further increase police presence, thereby perpetuating a vicious cycle.[130]

Incorporating algorithms and machine learning into the decisions that surround us has the potential to exacerbate inequities. Data analytics are sufficiently complex such that even the developers themselves often cannot determine why certain variables gave rise to certain outputs.[131] When opaque algorithms give rise to disparate treatments on the basis of race, gender, or other potentially sensitive variables—particularly in areas of importance—we should rightfully be concerned about the implications and seek out a framework, ethical or legal, that can appropriately inform or, when needed, constrain the decision-making that goes into the development and deployment of these tools.

5. Individuals Cannot Meaningfully Opt Out

The upshot of this novel, networked data landscape is that our behavior is constantly monitored and analyzed as we navigate our online and offline worlds. Data about our activities are collected, shared, aggregated, and analyzed and give rise to consequences that shape the world around us.[132] The systems that are established leave us little opportunity to protect ourselves from unjust consequences or unfair inferences—in other words, no meaningful opportunity to opt out.

It could be argued that our continued willingness to share data signals that we do not care much about privacy.[133] Researchers have coined the phrase "privacy paradox" to grapple with the ways that individuals express privacy

[128] Crawford, *supra* note 105.

[129] *Id.*

[130] *Id.* ("At the very least, this software risks perpetuating an already vicious cycle, in which the police increase their presence in the same places they are already policing (or overpolicing), thus ensuring that more arrests come from those areas. In the United States, this could result in more surveillance in traditionally poorer, nonwhite neighborhoods, while wealthy, whiter neighborhoods are scrutinized even less.").

[131] *See, e.g.*, Weissman, *supra* note 120.

[132] *Id.*

[133] Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 413 (2014) ("The problem is not that privacy is dead but rather that the system of managing the flows of personal information needs to be rethought in the face of the new uses and sources that our Information Revolution has generated.").

concerns but "rarely take action to stop cookies and other tools deployed to gather their data."[134] Researchers at the Massachusetts Institute of Technology found that even those who express concerns about privacy will share their friends' email addresses for free pizza.[135] And people are willing to reveal private details when we see others we trust doing so, or when we really need a mobile application or service that does not readily allow opt out.[136]

One response to the "privacy paradox" is not that individuals do not actually care about privacy but that the obstacles to exercising meaningful control over our data are simply too steep to overcome.[137] Behavioral economics research has shown that individuals tend to keep default settings, even with settings that could easily be changed.[138] In 2005, Facebook's default settings were a mix of sharing with friends and, at most, "friends of friends."[139] By 2010, the default had shifted to sharing with everyone.[140] Those who do not diligently monitor Facebook's changing default settings—and the shifting defaults across the entire digital landscape—will inadvertently share far more about themselves than they once had, and perhaps more than they ever intended to.

Researchers have also identified what they call the "paradox of control": The more that individuals believe they have control over their data, the more willing they are to share.[141] In a research experiment, individuals were asked ten personal questions: One group was told that if they answered, their responses would be publicly posted; a second group was allowed to choose not to publicly post their responses.[142] The appearance of control made those in the second group twice as likely to agree to share their responses publicly.[143] Facebook has learned this lesson. In his testimony before Congress, Mark Zuckerberg

[134] Porter, *supra* note 33; Eduardo Porter, *The Facebook Fallacy: Privacy Is Up to You*, N.Y. TIMES, Apr. 25, 2018, at B1; Idris Adjerid et al., *The Paradox of Wanting Privacy but Behaving as if it Didn't Matter*, LSE BUS. REV. (Apr. 19, 2018), http://blogs.lse.ac.uk/businessreview/2018/04/19/the-paradox-of-wanting-privacy-but-behaving-as-if-it-didnt-matter/ ("The term refers to apparent inconsistencies between people's stated privacy behavioural intentions and their actual behaviours.").

[135] Porter, *supra* note 134.

[136] Selyukh, *supra* note 21.

[137] Porter, *supra* note 134.

[138] *Id.*

[139] *Id.*

[140] *Id.*

[141] Cathy Cunningham, *Help Squad: Research Shows Greater Online Privacy Controls Can Result in Sharing More Personal Information*, CHI. TRIB. (May 17, 2018, 2:00 PM), https://www.chicagotribune.com/suburbs/ct-ahp-column-help-squad-tl-0524-story.html ("What happens with very granular controls is that people feel empowered, and at least in our research, the perception of control over personal information decreased privacy concerns.").

[142] *Id.*

[143] *Id.*

repeatedly emphasized how much control Facebook gives users over their personal data, enabling them to make their own privacy decisions.[144] Those familiar with "paradox of control" research understand the extent to which Facebook's response does not actually address the underlying concern.

Companies, too, have also made it harder for individuals to discern what is actually happening to their data. Researchers who analyzed Facebook's privacy policy on 33 different variables found that, between 2005 and 2015, Facebook's rating declined in 22 out of 33 measures of privacy protection and transparency.[145] The increasing inscrutability is by design.[146] The upshot is that, in practice, even vigilant users will have a more difficult time determining the uses to which their data has been put or the reasoning behind decisions made about their data or their lives.[147]

Finally, the reality is that people are imperfect decision-makers, particularly with choices that involve immediate gratification or delayed, but uncertain, potentially negative consequences.[148] In a world with changing defaults and lengthy, impenetrable privacy policies, individuals have little ability to opt out or engage in meaningful self-protection from data collection practices. Because of the ubiquity of data collection practices, the potential severity of the consequences, and the inability of individuals to meaningfully opt out, data collection must be subject to guidance. This guidance can come in the form of a legal or an ethical framework. As discussed in the next Parts, however, existing laws and ethical frameworks fall woefully short.

II. EXISTING LAWS PROVIDE INSUFFICIENT PROTECTION

In the current novel, networked data landscape, data about our online and offline activities are collected, aggregated, shared, and analyzed. This data collection and use gives rise to consequential decisions being made about us,

[144] Porter, *supra* note 134.

[145] Jennifer Shore & Jill Steinman, *Did You Really Agree to That? The Evolution of Facebook's Privacy Policy*, TECH. SCI. (Aug. 11, 2015), http://techscience.org/a/2015081102.

[146] Michelle De Mooy, *The Ethics of Design: Unintended (But Foreseeable) Consequences*, CTR. FOR DEMOCRACY & TECH. (Jan. 31, 2018), https://cdt.org/blog/the-ethics-of-design-unintended-but-foreseeable-consequences ("It is by design that it's nearly impossible for most people to know what's happening to their information in digital systems, no doubt because many people express intense discomfort when they learn how and why their data is used by companies.").

[147] *See id.*

[148] Cunningham, *supra* note 141 ("The problem is that while the benefits [of online behaviors] are always very visible to us, the costs are very much hidden. What does it mean in terms of our data privacy to actually do an action online? We don't know. The benefits are immediate and certain; the risks are only in the future and they are very much uncertain, so that makes our decision-making very, very hard.").

with limited ability to challenge the outcome and no meaningful ability to opt out or engage in self-protection. In such circumstances, the law could serve as an important backstop, regulating data activities and granting individuals meaningful rights in their data. Existing U.S. laws, however, fall woefully short. This Part considers the limitations of the sectoral approach to privacy in the United States, the shortcomings of even comprehensive data privacy laws being enacted and proposed, and the massive de-identification loophole in all data privacy laws that renders existing laws inadequate and leaves individuals under-protected. Ultimately, without meaningful legal guidance to constrain concerning data practices, ethical frameworks ought to guide data decision-making.

A. The Sectoral Nature of U.S. Privacy Law Offers Patchwork Protections

The networked nature of today's data landscape means that data are readily transferred across traditional boundaries. Whereas data were once contained within silos—credit card purchases by credit card companies, video rentals by video rental companies—data today are shared among disparate parties.[149] Data collected for one purpose—for example, a credit card company collecting purchase data from its customer—are now sold to unrelated, and often invisible, third parties.[150] Despite the ways that data today move seamlessly across boundaries, U.S. privacy law takes a sectoral approach to regulation.

Over the past five decades, the United States has enacted a series of laws designed to protect very specific types of personal data. The Fair Credit Reporting Act specifies when consumer credit information may be transmitted to third parties.[151] Educational records are protected under the Family Educational Rights and Privacy Act.[152] Video rental records are protected under the Video Privacy Protection Act of 1988.[153] Children under thirteen are afforded protections online under the Children's Online Privacy Protection Act.[154] Limits are placed on the ability of financial institutions to disclose private financial data under the Gramm-Leach-Bliley Act.[155] And a number of laws—

[149] Maréchal, *supra* note 32.

[150] *Id.*

[151] *See* Fair Credit Reporting Act, 15 U.S.C. § 1681 (2012).

[152] *See* Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g(b) (2012).

[153] *See* Video Privacy Protection Act of 1988, 18 U.S.C. § 2710(b) (2012).

[154] *See* Children's Online Privacy Protection Act, 15 U.S.C. § 6502(a)(1).

[155] *See* Gramm-Leach-Bliley Act, Pub. L. No. 106–102, tit. V, § 501(a), 113 Stat. 1338, 1436 (1999) (codified at 15 U.S.C. § 1601 (2012)).

primarily the Privacy Act of 1974[156]—offer protections when the U.S. government itself collects data.

Two laws—the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and the Federal Policy for the Protection of Human Subjects (the "Common Rule")[157]—are considered in depth below because they provide cross-cutting data privacy protections that are potentially applicable to this novel, networked data landscape, compared to the more narrow, sectoral laws sets forth above. As discussed below, however, even those protections are insufficient to address the concerns of networked data.

1. Health Insurance Portability and Accountability Act of 1996

HIPAA is one of the more robust forms of privacy protection in the United States, and offers protections for a subset of health data, a category that, along with financial data, is generally considered among the most sensitive in the nation. Although HIPAA is commonly understood to protect health information, the actual scope of HIPAA is far more limited. HIPAA governs the use and disclosure of "protected health information"—including name, address, date of birth, or identifiable information that relates to the past, present, or future medical care [158]—by "covered entities"—health care providers, health clearinghouses, health plans, or their business associates.[159]

HIPAA's Privacy Rule was finalized in 2002, before troves of health data were collected outside the healthcare infrastructure.[160] This means that categories of data commonly thought of as health data fall outside HIPAA's protections.[161] Health data collected by wearable fitness trackers, social media sites, and online health management tools fall outside HIPAA's protections.[162]

[156] *See* Privacy Act of 1974, 5 U.S.C. § 552a(b) (2012).

[157] *Federal Policy for the Protection of Human Subjects ('Common Rule')*, U.S. DEP'T OF HEALTH AND HUM. SERVS., https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html (last visited Nov. 27, 2019).

[158] Protected health information includes common identifiers such as name, address, birthday and Social Security number, or identifiable information that relates to the past, present, or future medical care. *See* U.S. DEP'T OF HEALTH & HUMAN SERVS., OCR PRIVACY BRIEF: SUMMARY OF THE HIPAA PRIVACY RULE 3–4 (2003) [hereinafter OCR PRIVACY BRIEF—HIPAA].

[159] *Id.* at 2–3.

[160] The Privacy Rule, finalized in 2002, sets limits on the disclosure of protected health information by covered entities without the consent of the data subject. *Id.* at 2.

[161] Angela Chen, *Why It's Time to Rethink the Laws that Keep Our Health Data Private*, VERGE (Jan. 29, 2019, 8:30 AM), https://www.theverge.com/2019/1/29/18197541/health-data-privacy-hipaa-policy-business-science ("Perhaps the biggest weakness of HIPAA, and the way that it underprotects [sic] us, is that it doesn't cover the enormous amount of data we generate in daily life that can hold clues to our health ….").

[162] U.S. DEP'T OF HEALTH & HUMAN SERVS., EXAMINING OVERSIGHT OF THE PRIVACY & SECURITY OF

Information provided by direct-to-consumer genetic testing companies also falls outside of HIPAA's protections.[163] Data about medications purchased using a drugstore loyalty card may be unprotected under HIPAA.[164] Data that can make health predictions—based on how much you spend at fast food restaurants, whether you have a gym membership, and how much television you watch[165]—fall outside HIPAA's protections as well.

The easy transmissibility of data, and the porousness of the boundary between health and non-health data, highlight the limitations of HIPAA's privacy protections when applied to the novel data landscape. For example, a genetic test conducted by a medical professional included in a medical record is protected by HIPAA. A direct-to-consumer genetic test with results transmitted directly to the user falls outside of HIPAA protections. If the consumer shares the genetic test results with a medical provider who includes the results in the patient's medical record, the results in the record would be covered by HIPAA.[166]

Core to the HIPAA regime is that protections do not apply to data that has been de-identified in accordance with HIPAA methodology.[167] Information that has been de-identified in accordance with HIPAA requirements can be shared and used without limit[168]—a loophole that, as discussed in Section III.C—becomes larger and less protective with time.

HEALTH DATA COLLECTED BY ENTITIES NOT REGULATED BY HIPAA 1, 32 (2016).

[163] Michael Schulson, *Spit and Take*, SLATE (Dec. 29, 2017, 12:04 PM), https://slate.com/technology/2017/12/direct-to-consumer-genetic-testing-has-tons-of-privacy-issues-why-is-the-industry-booming.html.

[164] David Lazarus, *CVS Thinks $50 Is Enough Reward for Giving Up Healthcare Privacy*, L.A. TIMES (Aug. 15, 2013, 12:00 AM), https://www.latimes.com/business/la-xpm-2013-aug-15-la-fi-lazarus-20130816-story.html.

[165] Joseph Jerome, *Where Are the Data Brokers?*, SLATE (Sept. 25, 2018, 7:30 AM), https://slate.com/technology/2018/09/data-brokers-senate-hearing-privacy.html.

[166] Chen, *supra* note 161 ("If you take an electrocardiogram (EKG) at the doctor, and the doctor puts the results into an electronic health record, that is protected by HIPAA because it's within the health care system. If you take an EKG with the Apple Watch and don't share that information with your doctor, that same information is not protected by HIPAA. But if you take an EKG using the new Apple Watch and share it with your doctor and she puts it in her electronic health records, it is protected by HIPAA.").

[167] The HIPAA Privacy Rule sets forth two methods of de-identification: (1) a formal determination by a qualified expert that the risks of re-identification by the intended recipient are quite small, or (2) a safe harbor that entails removal of eighteen specified identifiers. OCR PRIVACY BRIEF—HIPAA, *supra* note 158, at 4 ("There are no restrictions on the use or disclosure of de-identified health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual.").

[168] *See id.*; Chen, *supra* note 161 ("If you strip away information like name and Social Security number and picture, you're allowed to share the data without HIPAA restrictions ….").

2. *The Common Rule*

Data scientists and privacy professionals have increasingly turned to the Common Rule to guide data activities.[169] The Common Rule, initially published in 1991 with revisions that went into effect January 21, 2019,[170] offers important procedural protections for activities considered human subjects research. These protections include obtaining a data subject's informed consent to participate and prior third-party review by an institutional review board. Many of the data activities in today's networked data landscape will fall outside the definition of human subjects research and thus the important procedural protections offered by the law will be unavailable.[171]

Many of today's data activities will fall outside the Common Rule because they are not exactly research and do not quite involve human subjects.[172] Research is defined as "a systematic investigation … designed to develop or contribute to generalizable knowledge."[173] Large swaths of data activities, including most of the data activities discussed in Section II.A, fall outside this definition.

Moreover, only research using *identifiable* private information will be considered human subjects research; research using de-identified data falls outside the Common Rule requirements altogether.[174] The Common Rule leaves determinations about when information should be considered identifiable to consultation with appropriate experts, including those with expertise in data matching and re-identification.[175] The Common Rule therefore relies on a distinction between human subjects, who are afforded protections, and the data that humans give rise to, which is often unprotected. This distinction minimizes

[169] *See generally* Protection of Human Subjects, 45 C.F.R. § 46 (2010) [hereinafter Common Rule]; Danah Boyd & Jacob Metcalf, *Example "Big Data" Research Controversies*, COUNCIL FOR BIG DATA, ETHICS, & SOC'Y (Nov. 10, 2014), https://bdes.datasociety.net/wp-content/uploads/2016/10/ExampleControversies.pdf.

[170] *OHRP Revised Common Rule*, U.S. DEP'T HEALTH & HUM. SERVS., https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html (last visited Jan. 27, 2020).

[171] Metcalf et al., *supra* note 79, at 3 ("For many U.S. scholars in medicine, biology, and social science, the commitment to ethical research involving human subjects starts with an obligation to the ethical principles underpinning the Common Rule ….Yet, this rule is not designed for the type of work typically done under the purview of big data, raising significant questions for consideration.").

[172] Metcalf et al., *supra* note 80, at 8 ("[T]he large majority of data science … largely avoids these regulations by not quite qualifying as 'human subjects' because it does not involve an intervention in a subject's life, and not qualifying as 'research' because it does not collect new data in pursuit of generalizable knowledge.").

[173] Common Rule, *supra* note 168, § 46.102(k)(1).

[174] *Id.* § 46.102(e)(1)(ii).

[175] *Id.* § 46.102(e)(7)(i).

the ways that the humans from whom data is collected could nevertheless be harmed as a result of their data being used.[176]

The subset of data activities that may fall within the purview of the Common Rule may nevertheless be exempt from Common Rule requirements if the data being used are "publicly available."[177] The reasoning for this exemption is that "research methods using existing public datasets pose such miniscule risks to individual human subjects that researchers should not face scrutiny by IRBs."[178] Some have questioned whether this still holds true given that data analytics aggregate disparate datasets to unexpected—and deeply personal—effects.[179] As an example, data analytics can combine distinct publicly available datasets that all include data about a specific individual to make inferences and reveal information—about, for example, political views, sexual preferences, and immigration status—that may nevertheless not be public "in a colloquial sense because the subject has chosen to represent themselves partially and differently in various online spaces."[180]

Facebook's 2014 "emotional social contagion" study ignited debate among researchers about when data use should be subject to the Common Rule. Data scientists from Facebook and Cornell University published a study in the prestigious *Proceedings of the National Academies of Science* detailing the results of an experiment in which the Facebook feeds of nearly 700,000 people had been systematically modified.[181] The researchers found that the emotional valance of posts that show up in a user's news feed have emotional consequences for the user who sees them.[182] Facebook users whose feeds were modified, and thus became unwitting research participants, were never granted the protections of the Common Rule because Facebook claimed that the activities were not human subjects research such that the activities were instead governed by its terms of service, which permitted these types of modifications.[183]

[176] Matthew Zook et al., *Ten Simple Rules for Responsible Big Data Research*, 13 PLOS COMPUTATIONAL BIOLOGY 1 (2017) ("While the connection between individual datum and actual human beings can appear quite abstract, the scope, scale, and complexity of many forms of big data creates a rich ecosystem in which human participants and their communities are deeply embedded and susceptible to harm.").

[177] Common Rule, *supra* note 169, § 46.104(d)(4)(i).

[178] Metcalf et al., *supra* note 80, at 9.

[179] *Id.* (describing "data analytics techniques that can create a composite picture of persons from widely disparate datasets that may be innocuous on their own but produce deeply personal insights when combined").

[180] Jake Metcalf, *Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes*, COUNCIL FOR BIG DATA, ETHICS, & SOC'Y (Apr. 22, 2015), https://bdes.datasociety.net/wp-content/uploads/2016/10/Human-Subjects-Lit-Review.pdf.

[181] Metcalf et al., *supra* note 80, at 8.

[182] *Id.*

[183] *Id.*

Although the Common Rule offers important procedural protections—including informed consent and prior third-party review—it applies only to the data activities considered "human subjects research" that come within its ambit. Although the Common Rule potentially provides significant protections for data use in a wide range of subject matter areas, many of today's consequential data activities will not be covered.

B. Comprehensive Data Privacy Legislation Falls Short

The past several years has seen movement toward more comprehensive data privacy legislation. The European Union (EU) General Data Protection Regulation (GDPR) came into effect May 2018 and became the global standard bearer in data protection. Shortly thereafter, California followed suit with the California Consumer Privacy Act of 2018. Federal lawmakers are currently considering more comprehensive data privacy legislation. Although these laws provide important protections, they fall short of addressing the complex range of data concerns.

1. European Union General Data Protection Regulation

In May 2018, the European Union's sweeping data privacy law, the GDPR, came into effect. The GDPR recognized the protection of personal data as a fundamental right,[184] albeit one that must be balanced against other fundamental rights.[185] Because of the steep fines for noncompliance—up to twenty million Euros or 4% of total worldwide annual turnover—stakeholders around the world took notice and those who did business in the European Union took steps to comply.[186]

The GDPR itself affords individuals some meaningful protections to personal data. For example, it grants individuals rights to access,[187] rectify,[188] and erase data about themselves in certain circumstances.[189] It grants individuals the right to have automated decisions that are made about themselves reviewed

[184] Commission Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR].

[185] GDPR, *supra* note 184, 2016 O.J. (L 119) 2.

[186] GDPR, *supra* note 184, 2016 O.J. (L 119) 82.

[187] GDPR, *supra* note 184, 2016 O.J. (L 119) 43.

[188] GDPR, *supra* note 184, 2016 O.J. (L 119) 43.

[189] GDPR, *supra* note 184, 2016 O.J. (L 119) 43.

by humans[190] and the right to data portability.[191] It also gives data controllers responsibilities, like considering the privacy impacts of high-risk new technologies,[192] or appointing a Data Protection Officer.[193]

There are, however, significant ways that the GDPR falls short. First, although the GDPR generally requires a person's consent to data collection and use, the consent requirements can be satisfied with check-the-box consent (described approvingly as "ticking a box when visiting an internet website").[194] Notice and check-the-box consent are generally considered lacking as a meaningful consumer protection.[195]

Second, the protections of the GDPR only apply to data that are identifiable.[196] The GDPR takes a sweeping approach to determining identifiability taking into account "all the means reasonably likely to be used" to re-identify data including "all objective factors" such as the costs, amount of time, and technology available.[197] The GDPR also takes into account the ways that individuals may be tracked by "online identifiers" that "when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them."[198] As discussed in Section III.C, in an era of big data, however, when artifacts with limited connection to personal identity—including internet browser configuration or cell phone battery percentage—can track individuals across the Internet, this approach to identifiability is either too narrow to be protective or so all-encompassing as to swallow the original intent.[199]

[190] GDPR, *supra* note 184, 2016 O.J. (L 119) 46.

[191] GDPR, *supra* note 184, 2016 O.J. (L 119) 45.

[192] GDPR, *supra* note 184, 2016 O.J. (L 119) 53.

[193] GDPR, *supra* note 184, 2016 O.J. (L 119) 55.

[194] GDPR, *supra* note 184, 2016 O.J. (L 119) 6.

[195] De Mooy, *supra* note 146 ("Ethical products and services cannot rely on 'checking a box' for the use of customer data because such blanket consent ignores difficult questions about user expectations and the unique risks that data exposure might cause for one individual or group over another. Today's notice and consent mechanisms have become compliance tools for avoiding liability."); *see* Richards & King, *supra* note 13, at 412 ("[I]n practice most companies provide constructive notice at best, and individuals make take-it-or-leave-it decisions to provide consent."); Idris Adjerid et al., *supra* note 134.

[196] GDPR, *supra* note 184, 2016 O.J. (L 119) 33; *GDPR*, *supra* note 183, 2016 O.J. (L 119) 5 ("The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person.").

[197] GDPR, *supra* note 184, 2016 O.J. (L 119) 5.

[198] GDPR, *supra* note 184, 2016 O.J. (L 119) 6.

[199] Ohm, *supra* note 11, at 1704 ("Today, this debate centers almost entirely on squabbles over magical phrases like 'personally identifiable information' (PII) or 'personal data.' Advances in reidentification expose how thoroughly these phrases miss the point. Although it is true that a malicious adversary can use PII such as a name or social security number to link data to identity, as it turns out, the adversary can do the same thing using information that nobody would classify as personally identifiable.").

The third challenge is that the law itself cannot guide data decision-making. The law is long and complex enough that cottage industries have risen to advise companies about how to comply.[200] This complexity means that data scientists cannot internalize guiding principles and imbue them into their data activities. Even those most well-versed in GDPR requirements cannot turn to the GDPR for guidance about whether certain data practices *should* proceed—the GDPR establishes boxes that must be checked to ensure compliance with requirements but does not guide ethical decision-making.[201]

2. *The California Consumer Privacy Act of 2018*

In 2018, California decided to follow the European Union's lead in protecting personal data by enacting the California Consumer Privacy Act (CCPA).[202] The law's preamble recognized the changing data landscape,[203] including the ways that consumers are required to hand over information in exchange for goods and services.[204] The CCPA states that Californians retain a reasonable expectation of privacy in their personal information even when disclosed to a third party.[205]

The CCPA therefore offers consumers the right to: (1) know what personal information is being collected about them;[206] (2) know whether personal information is sold or disclosed and to whom;[207] (3) opt out of the sale of that information;[208] and (4) be offered equal service and price.[209] Companies must post a "clear and conspicuous" link on the company's website that allows a consumer to opt out of the sale of their personal information.[210]

[200] Salvador Rodriguez, *Business Booms for Privacy Experts as Landmark Data Law Looms*, REUTERS (Jan. 22, 2018, 7:10 AM), https://www.reuters.com/article/us-cyber-gdpr-consultants/business-booms-for-privacy-experts-as-landmark-data-law-looms-idUSKBN1FB1GP.

[201] Neil Hodge, *EU Regulator Pushes for Global Consensus on Data Ethics*, COMPLIANCE WEEK (Oct. 26, 2018, 2:16 PM), https://www.complianceweek.com/data-privacy/eu-regulator-pushes-for-global-consensus-on-data-ethics/2105.article ("The fact is that the European legislator did not think about ethics when it drafted the GDPR.").

[202] *See* California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.198(a) (West 2018).

[203] *See id.* pmbl. § 2(c).

[204] *See id.* ("It is almost impossible to apply for a job, raise a child, drive a car, or make an appointment without sharing your personal information."); *see also id.* pmbl. § 2(e).

[205] *See id.* pmbl. § 2(a)

[206] *Id.* pmbl. § (2)(i)(1).

[207] *Id.* pmbl. § (2)(i)(2).

[208] *Id.* pmbl. § (2)(i)(3).

[209] *Id.* pmbl. § (2(i)(5).

[210] *Id.* § 1798.135(a)(1).

Although the CCPA makes important advances, these consumer protections nevertheless fall short. First, the law protects only those instances where a consumer's personal data are sold and not when personal data are given away for free.[211] Should companies decide it is in their business interests to give personal data away for free, the company would be allowed to do so without limitation.[212] The CCPA, therefore, does not protect against decisions like the one Facebook made to give applications access to user data for free, which gave rise to the Cambridge Analytica breach.

Second, the CCPA requires disclosure only of the categories of information being collected or sold, rather than specific details.[213] Telling consumers that "commercial information" is collected and sold paints a different picture than making clear that what is being collected and sold to third parties is an itemized list of everything an individual has purchased.[214]

Third, the law itself does not protect de-identified data.[215] The CCPA applies to personal information that "identifies, relates to, describes, references, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or device."[216] As was true of the GDPR's approach, and as is discussed in more detail in Section III.C, this definition when implemented is either so broad as to render nearly all data identifiable or so narrow that large swaths of data unprotected.

Finally, the CCPA does not provide meaningful guidance to data decision-makers. The law tells companies what information they must provide consumers and requires that companies provide a mechanism for consumers to opt out of the sale of personal information, but does not offer guidance to companies about whether they *should* collect data or the factors they should weigh when deciding whether or not to sell consumers' data.

[211] This is a modification from the language in the original ballot initiative that defined "sale" as including sharing data for free. *See* Adam Schwartz et al., *How to Improve the California Consumer Privacy Act of 2018*, ELECTRONIC FRONTIER FOUND. (Aug. 8, 2018), https://www.eff.org/deeplinks/2018/08/how-improve-california-consumer-privacy-act-2018 ("By contrast, the original ballot initiative defined "sale" to include sharing data with other businesses for free.").

[212] *See* Schwartz et al., *supra* note 210.

[213] *See id.*; CIV. § 1798.100(a)–(b).

[214] *Id.* § 1798.140(0)(1)(D).

[215] *Id.* § 1798.145(a)(5).

[216] *Id.* § 1798.140(o)(1).

3. *Proposed Federal Privacy Legislation*

Despite decades of a sectoral approach to privacy in the United States, there is some movement toward more comprehensive federal privacy legislation.[217] The CCPA, enacted in 2018, took effect January 1, 2020. Companies now conceivably have to comply with the CCPA, to the extent they do business in California;[218] the EU's GDPR, to the extent a company engages users in the European Union; the sector-specific privacy laws, including HIPAA; and the patchwork of state privacy laws that already exist and that will be developed.[219]

The result is an emerging, and somewhat unexpected, consensus around enacting federal privacy legislation.[220] Amazon, Apple, and Google have supported the enactment of federal privacy legislation.[221] Federal privacy legislation has bipartisan support in Congress.[222] The Trump Administration,

[217] *See* David Meyer, *In Privacy We Trust*, FORTUNE, Dec. 1, 2018, at 38 ("A year ago, the idea of a federal data privacy law in the U.S. was unthinkable for all but a handful of digital rights activists. As 2018 comes to a close, the prospect of such legislation has suddenly become very real.").

[218] *See id.* at 39 ("Suddenly, tech firms were facing the prospect of disparate data privacy rules across different states. And that's when their calls for a comprehensive federal law began to coalesce."); Chen, *supra* note 160 ("To avoid getting in trouble, every institution needs to follow the policy of the state with the most restrictive laws. The result is that California, with its tough new data privacy law, is essentially setting the national policy.").

[219] Dan Clark, *Federal Data Privacy Legislation Is Likely Next Year, Tech Lawyers Say*, LAW.COM (Nov. 29, 2018, 5:00 PM), https://www.law.com/corpcounsel/2018/11/29/federal-data-privacy-legislation-is-likely-next-year-tech-lawyers-say/?slreturn=20190102162337; Mark Sullivan, *Inside the Upcoming Fight over a New Federal Privacy Law*, FAST CO. (Jan. 4, 2019), https://www.fastcompany.com/90288030/inside-the-upcoming-fight-over-a-new-federal-privacy-law.

[220] Despite apparent consensus in support of enacting legislation, there are likely to be disparities in approaches to the provisions contained in any proposal. *See* Issie Lapowsky, *Get Ready for a Privacy Law Showdown in 2019*, WIRED (Dec. 27, 2018, 7:00 AM), https://www.wired.com/story/privacy-law-showdown-congress-2019/ ("Parties on all sides of the privacy argument, for instance, say that people should be able to see what data is collected about them and how it's being shared. They also agree that companies should be required to get consent before processing user data, and that consumers should be able to request that their data be corrected or deleted. But there are [sic] a range of opinions on how those ideas should be implemented. Should companies be required to disclose every single piece of data they've collected on someone, or is sharing the categories of data enough? And what constitutes consent? Must consumers opt in to having their data processed, or is it sufficient to let them opt out?").

[221] *See* Meyer, *supra* note 217, at 38–39; Lapowsky, *supra* note 219 ("Companies like Amazon, Apple, Facebook and Google are pushing hard for federal digital privacy legislation in 2019, and not quite out of the goodness of their hearts …. [T]ech giants are racing the clock to supersede California's law with a more industry-friendly federal bill."); Robert Scammell, *US Tech Giants Back Federal Data Privacy Law, as Long as Innovation Is Protected*, VERDICT (Sept. 26. 2018, 7:25 PM), https://www.verdict.co.uk/us-tech-giants-federal-data-privacy-law/; Sullivan, *supra* note 219 ("What the tech industry wants is a federal privacy law that doesn't impose onerous or costly privacy requirements, and does not expand government powers to enforce the rules. And most important of all, the industry wants to make sure that a new federal law will supersede–or preempt–privacy laws enacted by the states, such as the tough privacy law passed by California, which is now scheduled to go into effect January 1, 2020.").

[222] *See* Scammell, *supra* note 221.

too, has announced that it looks to advance "a consumer privacy protection policy that is the appropriate balance between privacy and prosperity."[223]

The upshot is that several major pieces of federal privacy legislation have been introduced in the Senate.[224] A leading proposal, the Consumer Data Protection Act sponsored by Senator Ron Wyden (D-OR),[225] enhances the power of the Federal Trade Commission (FTC) to establish and enforce minimum privacy and security standards and fine companies for first offences, and authorizes the creation of a new Bureau of Technology within the FTC.[226] The proposal creates a universal "Do Not Track" option that allows individuals to opt out of third parties tracking, sharing, storing, and using their data; this option is retrospective, forcing companies that have already collected information about individuals who opt out to delete that data upon opt out.[227] The proposal has strict penalties for noncompliance, including up to 4% of total gross revenue and criminal penalties for CEOs and other executives.[228] The proposal is limited to companies that collect data on more than one million individuals and have annual revenue exceeding fifty million dollars.[229] The proposal does not require consent for data to be collected, so companies are still permitted to collect unlimited amounts of data about individuals.[230]

Additional proposals—including the CONSENT Act,[231] the Social Media Privacy Protection and Consumer Rights Act of 2018,[232] and the Information Transparency & Personal Data Control Act[233]—generally coalesce around

[223] Meyer, *supra* note 217, at 38; *see* Lapowsky, *supra* note 219 ("The Trump administration's National Telecommunications and Information Administration has released its own point-by-point proposal, describing in unspecific terms a set of 'privacy outcomes' the administration would like to see. It too proposes a bill that would 'harmonize the regulatory landscape' to 'avoid duplicative and contradictory privacy-related obligations.'").

[224] Sullivan, *supra* note 219.

[225] Ron Wyden, *The Consumer Data Protection Act of 2018 Discussion Draft*, RON WYDEN (Nov. 1, 2018), https://www.wyden.senate.gov/imo/media/doc/Wyden%20Privacy%20Bill%20one%20pager%20Nov%201.pdf.

[226] *Id.*; *see also* Allie Bohm, *How Well Do the Current Federal Privacy Proposals Protect Your Privacy?*, PUB. KNOWLEDGE (Dec. 21, 2018), https://www.publicknowledge.org/how-well-do-the-current-federal-privacy-proposals-protect-your-privacy/; Sarah Parker, *A Step Forward for Federal Privacy Legislation*, HARV. J.L. & TECH. DIG. (Dec. 5, 2018), http://jolt.law.harvard.edu/digest/a-step-forward-for-federal-privacy-legislation.

[227] Companies would also be required to query the list proactively before collecting data. *See* Wyden, *supra* note 225.

[228] *See id.*

[229] Parker, *supra* note 226.

[230] Bohm, *supra* note 226.

[231] CONSENT Act, S. 2639, 115th Cong. § 2 (2018).

[232] Social Media Privacy Protection and Consumer Rights Act of 2018, S. 2728, 115th Cong. § 2 (2018).

[233] Information Transparency & Personal Data Control Act, H.R. 6864, 115th Cong. § 2 (2018).

similar themes: expanding the reach of the FTC, requiring better notice to consumers, and allowing some measure of individual opt out. One proposal, the Data Care Act of 2018—introduced by Senator Brian Schatz (D-HI) and fourteen Democratic co-sponsors—instead introduces the concept of a data fiduciary with duties of care, loyalty, and confidentiality comparable to those already imposed on lawyers and doctors.[234] And still other proposals may emerge. For example, Senator Cory Booker (D-NJ) has focused on algorithmic transparency and social justice; those ideas may be incorporated in a future consensus bill.[235] All proposals thus far limit coverage to identifiable data.[236]

C. Exempting De-identified Data is Insufficiently Protective

As discussed, data that are de-identified fall outside existing privacy protections altogether.[237] De-identified data are exempted from coverage under HIPAA,[238] the Common Rule,[239] the GDPR,[240] the CCPA,[241] and proposed federal privacy laws.[242] Allowing the free exchange of de-identified data has been justified on the grounds that because the data do not identify individuals, sharing is essentially a risk-free endeavor.[243]

The reality of this novel, networked data landscape, however, is that datasets that were previously thought to be de-identified are now readily—and trivially—re-identifiable.[244] In fact, "as these datasets have proliferated, so too has research

[234] *See* Data Care Act of 2018, S. 3744, 115th Cong. § 2 (2018).

[235] Sullivan, *supra* note 219.

[236] *See* H.R. 6864 (defining "sensitive personal information" as "information relating to an identified or identifiable individual"); S. 3744 (governing "individual identifying data"); S. 2728 (defining "personal data" as "individually identifiable information about an individual collected online"); S. 2639 (governing "personally identifiable information").

[237] Ohm, *supra* note 111, at 1740 ("[A]lmost every single privacy statute and regulation ever written in the U.S. and the EU embraces—implicitly or explicitly, pervasively or only incidentally—the assumption that anonymization protects privacy, most often by extending safe harbors from penalty to those who anonymize their data.").

[238] OCR PRIVACY BRIEF—HIPAA, *supra* note 158 ("There are no restrictions on the use or disclosure of de-identified health information.").

[239] *Common Rule*, *supra* note 168, at § 46.102(e)(1)(ii) (defining human subjects research as the use of "identifiable private information or identifiable biospecimens").

[240] GDPR, *supra* note 184, 2016 O.J. (L 119) Recital 26 ("The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person ….").

[241] California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.140(h) (West 2018).

[242] *See* Part III.B.1.

[243] *See* ARVIND NARAYANAN ET AL., A PRECAUTIONARY APPROACH TO BIG DATA PRIVACY 13 (2015).

[244] *See id.* at 2; John Nosta, *Healthcare Data as Property Can Change Everything*, FORBES (June 5, 2018, 8:44 AM), https://www.forbes.com/sites/johnnosta/2018/06/05/healthcare-data-as-property-can-change-everything/ ("[D]ata isn't really de-identified. In fact, given the right constellation of a just few data points, re-

demonstrating that even the most carefully anonymized datasets can be de-identified with relative ease."[245] As the number of data points collected, aggregated, and shared online grows, so too does the ability to re-identify data subjects.[246]

The concept of "mosaicking"—or combining multiple innocuous datasets together to fill in the gaps in each[247]—explains why the ease of re-identification grows as more data are released.[248] A team at Columbia University collected declassified documents authored by several federal U.S. government agencies. Each agency took a different approach to declassifying the document—one agency might have redacted the first paragraph and released the rest, another agency might have redacted people's names, and another might have redacted the last paragraph and released the rest. By pooling multiple redacted versions, along with publicly accessible media reports, it was possible to develop a fairly complete picture of the declassified document.[249]

Time and again, "anonymized" datasets are released only to be trivially re-identified. The New York City Taxi & Limousine Commission released an "anonymized" dataset of over 1.1 billion individual taxi rides taken between 2009 and 2015.[250] A graduate student searched the Internet for pictures of "celebrities in taxis in Manhattan in 2013," and was able to see the taxi's medallion number in the picture, cross-reference the time and location against the taxi database, and identify the celebrity's destination and amount tipped.[251]

In 2006, AOL decided to make an "anonymized" dataset of 21 million searches conducted by 650,000 users available to researchers.[252] Names were removed from the search database and replaced by numeric identifiers.[253] It did

identification can become fairly simple, and even commonplace.").

[245] Kalev Leetaru, *The Big Data Era of Mosaicked Deidentification: Can We Anonymize Data Anymore?*, FORBES (Aug. 24, 2016), https://www.forbes.com/sites/kalevleetaru/2016/08/24/the-big-data-era-of-mosaicked-deidentification-can-we-anonymize-data-anymore/.

[246] *See* NARAYANAN ET AL., *supra* note 243, at 2 (quoting PRESIDENT'S COUNCIL OF ADVISORS ON SCI. AND TECH., REPORT TO THE PRESIDENT: BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE 38–39 (2014)).

[247] *See* Leetaru, *supra* note 245.

[248] *See* Ohm, *supra* note 11, at 1705 ("Reidentification combines datasets that were meant to be kept apart, and in doing so, gains power through accretion: Every successful reidentification, even one that reveals seemingly nonsensitive data like movie ratings, abets future reidentification. Accretive reidentification makes all of our secrets fundamentally easier to discover and reveal.").

[249] *See* Leetaru, *supra* note 245.

[250] *See id.*

[251] *Id.*

[252] Garfinkel & Theofanos, *supra* note 26, at 21; Leetaru, *supra* note 246.

[253] Garfinkel & Theofanos, *supra* note 26, at 21; Leetaru, *supra* note 246.

not take long for identities to become associated with the various searches because of how much personal information people reveal in their searches: "From vanity searches on one's name to checking the local weather to searches for parts for a particular automobile to how to treat a particular medical condition, just knowing the set of searches performed by a particular person can be used to fairly quickly re-identify that person."[254]

In October 2006, Netflix launched a challenge offering a prize of $1 million to anyone who could develop an algorithm better at predicting movie preferences than Netflix's existing predictive model.[255] To assist developers, Netflix released anonymized records from nearly 500,000 Netflix customers that included rental data and the customer's rating.[256] Graduate students from the University of Texas were able to correlate video ratings in the Netflix database with publicly available ratings on IMDb, which included names and other identifiers, to re-identify the previously de-identified users.[257]

The specificity and granularity of data collected about us, in connection with the vast troves of data already released, makes it easier to identify individuals in large, anonymized datasets. In late 2018, the *New York Times* released an exposé about the extent to which de-identified location data can nevertheless identify individuals.[258] The article included examples of cell phone GPS coordinates that traveled back and forth between the New York governor's mansion and the YMCA where the governor was known to exercise; phones traveling from a specific home address to a known work address, with stops at doctors' offices along the way; even signals sent from the Capitol steps during the inauguration indicating the location of President Trump and his associates.[259]

Databases can be re-identified even when no traditional identifiers are present. Researchers were successfully able to connect anonymized Internet browsing histories to Twitter profiles in about 70% of users.[260] Researchers were able to re-identify 94% of Airbnb hosts by cross-referencing against voter registration databases despite Airbnb attempting to keep identities private.[261]

[254] Leetaru, *supra* note 245.

[255] Garfinkel & Theofanos, *supra* note 26, at 21.

[256] *See id.*

[257] *Id.*

[258] *See* Valentino-Devries et al., *supra* note 12.

[259] *Id.*; Barry Devlin, *The Anonymization Myth*, TDWI: UPSIDE (Sept. 4, 2018), https://tdwi.org/articles/2018/09/04/dwt-all-anonymization-myth.aspx ("Recording geolocation metadata over time (from fitness wearables or smartphones, for example) produces unique, repeatable data patterns that can be directly associated with individuals.").

[260] Anderson, *supra* note 29; Su et al., *supra* note 29.

[261] Aron Szanto & Neel Mehta, *A Host of Troubles: Re-Identifying Airbnb Hosts Using Public Data*, TECH.

Anonymized databases of credit card transactions scrubbed of all PII can identify 90% of individuals with just the data and location of four transactions.[262]

Even health data—generally thought to be among the most protected under HIPAA and considered among the most sensitive—are increasingly capable of being re-identified. Dr. Latanya Sweeney—whose groundbreaking work strongly influenced the development of HIPAA's Privacy Rule[263]—has recognized the ways that existing approaches to de-identification are inadequate given the novel, networked data landscape.[264] Dr. Sweeney has been able to re-identify 25% of data subjects in a data set de-identified to HIPAA standards.[265]

In another project, Dr. Sweeney was able to put names to patient records 43% of the time by cross-referencing publicly available, de-identified state-wide records of hospitalizations against newspaper stories that contained the word "hospitalized."[266] As described by Dr. Sweeney:

> At first glance, linking patients to publicly released health database records may seem academic or simply a matter of curiosity. But having an ability to access the records allows employers to potentially check on employees' health, financial institutions to adjust credit-worthiness based on medical information, data-mining companies to construct personal medical dossiers, newspapers to uncover health information on public figures, and people to snoop on friends, family, and neighbors.[267]

The upshot is that by constructing legal systems that turn on whether data are identifiable or de-identified—true of every existing and proposed piece of privacy legislation—personal data are insufficiently protected.[268] The governing

SCI. (Oct. 9, 2018), https://techscience.org/a/2018100902/.

[262] Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCI. 536, 536 (2015); Scott Berinato, *There's No Such Thing as Anonymous Data*, HARV. BUS. REV. (Feb. 9, 2015), https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data.

[263] *See* Leetaru, *supra* note 245.

[264] *See* Latanya Sweeney et al., *Re-Identification Risks in HIPAA Safe Harbor Data: A Study of Data from One Environmental Health Study*, TECH. SCI. (Aug. 28, 2017), https://techscience.org/a/2017082801/ ("The HIPAA Safe Harbor is not sufficient to protect data against re-identification.").

[265] *See id.*

[266] Latanya Sweeney, *Only You, Your Doctor, and Many Others May Know*, TECH. SCI. (Sept. 29, 2015), http://techscience.org/a/2015092903.

[267] *Id.*

[268] Ohm, *supra* note 11, at 1704 ("Today, this debate centers almost entirely on squabbles over magical phrases like 'personally identifiable information' (PII) or 'personal data.' Advances in reidentification expose how thoroughly these phrases miss the point. Although it is true that a malicious adversary can use PII such as a name or social security number to link data to identity, as it turns out, the adversary can do the same thing using information that nobody would classify as personally identifiable."); Berinato, *supra* note 261 ("Broadly, it means that anonymity doesn't ensure privacy, which could render toothless many of the world's laws and

premise of U.S. privacy law—that removing specific identifiers protects against re-identification and thus can be shared without risk to the data subject—is demonstrably false.[269] With de-identified data wholly unprotected and identifiable data only protected to the extent they fall into narrow, sectoral privacy laws, the existing legal framework offers inadequate protections.

Data decision-making, which holds out so much potential for benefit but also the possibility of harm, is therefore not meaningfully constrained by current or proposed laws. To help ensure that the power of data analytics is used more for good than for harm, data decision-makers ought to be guided by an ethical framework. This next Part analyzes existing ethical frameworks and the ways that they fall short in meaningfully guiding data decision-making.

III. EXISTING ETHICAL FRAMEWORKS PROVIDE INSUFFICIENT GUIDANCE

As discussed in this Article, the novel, networked data landscape holds out tremendous potential for benefit. If data decisions are made improperly, there is potential for serious societal and individual harm. As discussed in Part III, existing and proposed laws are insufficient to meaningfully protect against the worst data abuses or to guide data decision-makers.

Given the limitations of existing and proposed privacy laws, some have turned instead to existing ethical frameworks to guide data decision-making.[270] Practitioners across the data landscape are at a stage that biomedical researchers once were: choosing to use people—then, their bodies; now, their data—as a means to a greater good—then, generalizable knowledge; now, the benefits of big data.[271] When abuses in the realm of biomedical research came to light, those across the biomedical landscape decided an ethical framework—the Belmont Report—would be needed to safeguard against future abuse and steer the industry to more ethical practices.[272] As abuses across the data landscape increasingly come to light, similar calls for ethical guidance are emerging.

regulations around consumer privacy.”).

[269] *See* de Montjoye et al., *supra* note 262, at 539.

[270] *See* Metcalf, *supra* note 180.

[271] *See* Leetaru, *supra* note 48 (“Much of our modern ethical infrastructure exists because the medical profession once chose the same path that our digital disciples are taking today. In the name of the greater good of society, it was once deemed acceptable to experiment on innocent individuals without their knowledge or consent, without their ability to opt out, without them having any control over their personal information or its sharing.”).

[272] Elizabeth Pike, *Recovering from Research: A No-Fault Proposal to Compensate Injured Research Participants*, 38 AM. J. L. & MED. 7, 15–16 (2012).

Existing ethical frameworks—the Belmont Principles and the Fair Information Practice Principles (FIPPs)—were designed to address the ethical concerns of human subjects research and discrete information collections, respectively. As discussed below, because of the emerging and evolving concerns raised by networked datasets, these existing ethical frameworks fall short. This Part articulates ways that several of these established ethical principles ought to be expanded or modified to address today's concerns. These proposed modifications are incorporated into the CRAFT framework, set forth in Section V.B.

A. Belmont Report

The Belmont Report came into existence in the wake of revelations about the Tuskegee syphilis study and egregious violations of human research subjects.[273] The *National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research* was tasked with "identify[ing] the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects."[274] In 1979, The *National Commission* published the Belmont Report, "a statement of basic ethical principles" to "assist in resolving the ethical problems" of human subjects research.[275] The National Commission's concrete proposals—including informed consent to participation and prior third party review by institutional review boards—became enshrined in U.S. law in the Common Rule.[276]

The Belmont Report articulates three ethical principles that guide human subjects research: (1) respect for persons, (2) beneficence, and (3) justice.[277] Respect for persons recognizes that "individuals should be treated as autonomous agents" who are "capable of deliberation about personal goals and of acting under the direction of such deliberation."[278] Respecting autonomy requires giving "weight to autonomous persons' considered opinions and choices while refraining from obstructing their actions unless they are clearly detrimental to others."[279] Application of respect for persons is carried out through informed consent: Investigators should provide information that

[273] *See Research Implications: How Tuskegee Changed Research Practices*, CTRS. FOR DISEASE CONTROL & PREVENTION, http://www.cdc.gov/tuskegee/after.htm (last updated Dec. 14, 2015).

[274] NAT'L COMM'N PROT. HUMAN SUBJECTS OF BIOMEDICAL & BEHAVIORAL RESEARCH, THE BELMONT REPORT 1 (1979) [hereinafter BELMONT REPORT].

[275] *Id.*

[276] *See* Subpart A—Basic HHS Policy for Protection of Human Research Subjects, 45 C.F.R. § 46 (2010).

[277]

[278] BELMONT REPORT, *supra* note 243, at Part B.

[279] *Id.*

reasonable persons would wish to know, investigators must be sure that subjects comprehend, and subjects must give voluntary agreement.[280]

Informed consent in this data landscape is too onerous if implemented fully and meaningless if not. Research informed consent documents are long, detailed, and filled with medical and legal information about the research protocol and the potential implications of a decision to participate—they are time-consuming to review, but manageable for a one-time decision to enroll in human subjects research.[281] Implementing meaningful informed consent for every transaction that gives rise to data is too onerous given the thousands of actions that give rise to data every day.[282] Consent as currently implemented means that data-collecting entities generally disclose everything in a long and detailed—sometimes vague and opaque—terms of service and require that users click "yes" to continue.[283] Research has shown how few people click through to the terms of service—fewer than one in a thousand—and of those who do, how few actually read the language: Among those who click, the median time spent reviewing is twenty-nine seconds.[284] Check-the-box consent does not foster customer appreciation for the implications of an agreement to share data.

Importantly, informed consent in the biomedical research context is predicated on a researcher who can know and articulate the risks and benefits of participating in research, and individuals who can appreciate and make considered judgments in response.[285] When an action becomes data, no one person or entity knows or can predict the places that the data could travel or the consequences that could result. As privacy scholar and professor Dr. Zeynep Tufekci has stated:

> In the digital age, there is NO meaningful informed consent with regards to data privacy that operates at an individual level. Current

[280] *Id.* at Part C(1).

[281] David B. Resnik, *Do Informed Consent Documents Matter?*, 30 CONTEMP. CLINICAL TRIALS 114, 114 (2009).

[282] Metcalf, *supra* note 180, at 9 ("Medical ethicists have noted that the emphasis on patient consent in medical practice both empowered individuals to more vocally express their preferences and burdened them with the responsibility for balancing complex measures of harm and benefit ….").

[283] *See, e.g.*, Andy Greenberg, *Who Reads the Fine Print Online? Less than One Person in 1000*, FORBES (Apr. 8, 2010, 3:15 PM), https://www.forbes.com/sites/firewall/2010/04/08/who-reads-the-fine-print-online-less-than-one-person-in-1000/.

[284] *Id.*

[285] Metcalf et al., *supra* note 80, at 7 ("As it becomes cheaper to collect, store, and re-analyze large datasets, it has become clear that informed consent at the beginning of research cannot adequately capture the possible benefits and (potentially unknown) risks of consenting to the uses of one's data.").

> implications are unknown; future uses are unknowable. Companies structurally cannot inform and we are in no position to consent.[286]

The ethical protection required by big data, therefore, is not respect for considered judgment—too onerous a requirement given the thousands of daily interactions that give rise to collectible data and too impracticable given the unknowns—but respect for individual choice writ large when transactions involve consequential data practices that are not reasonably foreseeable given the circumstances.[287] As described in Part V.B, modifications to the principle of informed consent are reflected in the first proposed principle from the CRAFT framework, Choice.

The second ethical principle from the Belmont Report, beneficence, requires that researchers do no harm, maximize possible benefits, and minimize possible harms.[288] As applied in human subjects research, an institutional review board conducts a "systematic, nonarbitrary analysis of risks and benefits" to determine "whether the risks that will be presented to the subjects are justified."[289]

In the world of big data, risks and benefits are not only unknowable and unquantifiable, but they also evolve over time.[290] What are the risks of disclosing an individual's data of birth? The risks depend on a seemingly infinite array of unknowable particulars.[291] Likewise, the benefits of data use may be difficult to quantify. What are the benefits of a search engine providing more fine-grained results? How would they be measured and to whom must they redound? Many of today's foundational technologies—Google, Facebook, Amazon—had unknowable or unquantifiable risk-benefit ratios at the outset, and likely at various intervals since.[292] Given the near impossibility of assessing the beneficence of data decisions, other ethical principles must serve as bulwarks.[293] Because of the impracticability of analyzing beneficence across the data landscape, the proposed CRAFT framework focuses on other ethical principles—Responsibility, Accountability, and Fairness—to ensure that data

[286] Romero, *supra* note 95 (quoting Zeynep Tufekci (@zeynep), TWITTER (Jan. 19, 2018, 9:31 AM), https://twitter.com/zeynep/status/957969336932630528?lang=en).

[287] *See* Part V.B.

[288] BELMONT REPORT, *supra* note 274, at Part B(2).

[289] *Id.* at Part C(2).

[290] *See* Cunningham, *supra* note 141.

[291] *See* Porter, *supra* note 33.

[292] *See id.* ("Unfortunately, we have only rudimentary tools to measure the good and the bad."); *see also* Porter, *supra* note 134 ("Could we face higher prices online because Amazon has a precise grasp of our price sensitivities? Might our online identity discourage banks from giving us a loan? What else could happen? How does the risk stack up against the value of a targeted ad, or a friend's birthday reminder?").

[293] *See infra* Part V.B.

decisions are made in consideration of the well-being of the data landscape, and individuals contained within it.

The third ethical principle, justice, reflects notions of distributive justice and requires the equitable distribution of benefits and burdens.[294] As applied, distributive justice is carried out through fair subject selection—ensuring the people are not being selected to participate in research "because of their easy availability, their compromised position, or their manipulability, rather than for reasons directly related to the problem being studied."[295] Importantly, although distributive justice requires fairness across society, it does not require fairness to any particular individual or procedural fairness. For that reason, the CRAFT framework proposes using a more expansive definition of fairness so as to bring into consideration societal, individual, and procedural fairness.

In a world where data are collected from and about everyone, fair subject selection offers limited protection. More importantly, the concerns implicated by novel, networked data landscape implicate broader concerns about justice than mere distributive justice. Rather, they require addressing concerns of both fairness to society—protecting against data uses that encode and exacerbate existing population-wide inequities—and fairness to individuals, ensuring that data subjects receive the fair transaction they reasonably expect to have entered into.[296]

B. *Fair Information Practice Principles*

The Fair Information Practice Principles (FIPPs) is an ethical framework designed to govern the collection, use, and sharing of personally identifiable information. The FIPPs—promulgated in 1973 by an advisory committee of the U.S. Department of Health, Education, and Welfare—establishes ethical principles subsequently enshrined in privacy laws around the globe, including the Privacy Act of 1974, sector-specific regulations, and the GDPR.[297]

The FIPPs, as articulated by the Federal Trade Commission, include: (1) notice, requiring that practitioners give notice of data uses before collection; (2) choice, granting individuals the right to opt in or out; (3) access, granting

[294] BELMONT REPORT, *supra* note 274, at Part B(3).

[295] *Id.*; *see also id.* at Part C(3).

[296] *See infra* Part V.B.

[297] Ohm, *supra* note 11, at 1733–34 ("Spurred by this, in 1973 an advisory committee created by the secretary of health, education, and welfare issued a report that proposed a new framework called 'Fair Information Principles' (FIPS). The FIPS have been enormously influential, inspiring statutes, law review articles, and multiple refinements.").

individuals the right to view data collected about them and verify or contest its accuracy; (4) integrity, keeping data accurate and secure; and (5) enforcement, ensuring the principles are complied with.[298] Other articulations, including by the Department of Homeland Security, include: (6) purpose specification, whereby practitioners articulate the purpose for data collection; (7) data minimization, whereby practitioners collect only the minimum amount of data; and (8) use limitation, ensuring data are used only for the purpose specified.[299]

Although foundational and important, these principles are inadequate to deal with the sprawling nature of networked datasets. The FIPPs were developed in a world where individuals interacted directly with a data collector and could consent, opt out, or personally hold accountable data collectors who overstepped their bounds. The principles envision data collectors who will specify a purpose of the data collection at the outset, collect only the minimum necessary, and use data only for purposes consistent with the initial consent. We are no longer in that world.

Accordingly, the FIPPs fall short of providing adequate ethical protection and are impracticable in the current data landscape. The first two principles—notice and choice—are the cornerstone of today's data landscape. Notice, as implemented, often means that users are provided with a long, dense description of vague, potential future uses designed to discharge legal duties rather than to facilitate understanding. Choice, as currently implemented, allows users to opt in or out of data collection, often with a concomitant denial of access to the service upon opting out. While notice and choice are integral to any data ethics framework, as conceived of in the FIPPs and as implemented today, they are altogether inadequate. The principle of notice is encapsulated and expanded upon in the CRAFT framework in the principle of Transparency—an ethos that is expected to guide all aspects of data decision-making. Choice is retained as a centerpiece of the CRAFT framework but in a way that is intended to be more encompassing than opt in or out.

The second set of FIPPs principles, access and integrity, do not address concerns at the heart of today's data landscape. Both access and integrity are aimed at ensuring accuracy of the data. In today's landscape, however, the concern is more often that the volume of data collected means that even the most accurate of collections can inadvertently reveal too detailed a picture of a data subject. In fact, some privacy scholars advocate specifically injecting artificial

[298] Fred H. Cate, *The Failure of Fair Information Practice Principles, in* CONSUMER PROTECTION IN THE AGE OF THE INFORMATION ECONOMY 352–53 (Jane K. Winn ed., 2006).

[299] U.S. DEP'T OF HOMELAND SEC., NO. 2008-01, PRIVACY POLICY GUIDANCE MEMORANDUM, (2008).

misinformation into the data landscape to make any determinations or inferences less precise.[300] Moreover, in an era where it was clear who was collecting data, rights to access data may have offered meaningful protection. However, in today's networked data landscape, individuals cannot know all entities that have data about them and thus are in no position to use access as a meaningful measure of self-protection. For that reason, the CRAFT framework focuses on the networked data flows and interrelationships of data holders and making the data collection, transmission, and overall use more ethical.

The third cohort—purpose specification, data minimization, and use limitation—all presume that a specific purpose for collecting data is known at the outset and can be articulated, that the entity collecting the data is the same as the entity using it, and that the data user is in a position to communicate directly to the data subject. The networked data landscape means these are not always the case. Moreover, these principles may no longer be desirable goals. Many of the potential benefits of data analytics arise from harnessing the power of data in unexpected ways. Limiting uses to those described at the outset, and collecting only the minimum necessary, may hamper important advances. For this reason, rather than limiting the collection and use of data, the CRAFT framework instead focuses on making the entire data decision-making process more ethical.

Finally, although the FIPPs articulate principles of data management, they do not provide clear guidance about when data collection, use, or sharing can ethically proceed. If a data collector minimizes the data collected, gives individuals the opportunity to rectify incorrect data points, and satisfies a few additional requirements, all FIPPs could conceivably be satisfied even if the reasons motivating the collection, use, and sharing are unethical.[301] For that reason, the next Part proposes a novel ethical framework—the CRAFT framework—capable of meaningfully guiding data decision-making.

[300] *See* Daniel C. Howe & Helen Nissenbaum, *TrackMeNot: Resisting Surveilliance in Web Search*, *in* LESSONS FROM THE IDENTITY TRAIL: ANONYMITY, PRIVACY AND IDENTITY IN A NETWORKED SOCIETY 1–5 (Ian Kerr, Carole Lucock & Valerie Steeves eds., 2009); *Clicking Ads So You Don't Have To*, ADNAUSEAM, https://adnauseam.io (last visited Nov. 29, 2019).

[301] For consideration of substantive limits on data collection, use, and disclosure, see Roger Allan Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1 (2016).

IV. THE PATH TO ETHICAL DATA GOVERNANCE

If this novel, networked data landscape is to give rise to its most promising benefits, data decision-makers must be provided with meaningful guidance that can help foster an ethical data landscape and protect against the worst data abuses. As discussed in Part III, existing and proposed laws are insufficient to meaningfully protect against abuse. More importantly, these laws do not offer decision-makers guidance about whether things that are legally permissible nevertheless hold out the prospect of more harm than good and thus should be reconsidered. Part IV considered existing ethical frameworks potentially applicable to this data landscape—the Belmont Report and FIPPs—and concluded that, although much can be learned from these approaches, they are generally ill-suited for this complex, networked data landscape. Nevertheless, the data landscape is in need of ethical guidance. In the wake of several high-profile data events, today's data landscape is at an inflection point.[302] Apple CEO Tim Cook has observed that society "will never achieve technology's full potential without the full faith and confidence of the people who use it."[303]

Full faith and confidence must be earned. It requires that individuals have trust in the ecosystem and in its decision-makers. This Part argues for the importance of data ethics in maintaining the public's full faith and confidence and proposes an ethical framework—the CRAFT framework—that builds upon the lessons learned from existing ethics frameworks and consideration of the data ethics literature writ large. This Part discusses ways that the CRAFT framework could be implemented to enact meaningful reform of the data ethics landscape.

A. *The Importance of Ethics*

In response to a series of high-profile data events that have exposed the extent to which people are being tracked, monitored, and possibly manipulated by technology, calls for a more ethical data landscape have grown louder.[304]

[302] *See* Randy Bean, *A Rising Crescendo Demands Data Ethics and Data Responsibility*, FORBES (Oct. 29, 2018, 6:32 PM), https://www.forbes.com/sites/ciocentral/2018/10/29/a-rising-crescendo-demands-data-ethics-and-data-responsibility/#6b511f0bb5d5 ("The increased focus and concern for the ethical use of data is born out of widespread reaction to recent and highly publicized misuses of data that represent breaches of public trust—whether this be unauthorized data sharing by social media platforms, reselling of customer information by businesses, or biased algorithms that reinforce social inequalities.").

[303] Meyer, *supra* note 217, at 38.

[304] Bean, *supra* note 302 ("A spate of recent articles … underscore the increasing urgency and highlight the ethical considerations that organizations must address when managing data as an asset, and considering its impact on individual rights and privacy.").

Although discussions are currently underway about federal privacy legislation,[305] laws are limited in their ability to address emerging and evolving concerns.[306] The legislative process is, by its nature, a deliberative one, and it may be hampered in its ability to address these fast-moving, innovative technologies.[307] It may also be premature to invoke the full force of law in a nascent field due to the risk of shutting down promising avenues of inquiry.[308]

We need not, however, choose between legislation and ethical guidance.[309] Robust ethical guidance can inform comprehensive legislation. Just as the Belmont principles became enshrined in the Common Rule, and the FIPPs provide the framework for privacy laws around the world, so too could a data ethics framework guide comprehensive data legislation. As the novel, networked data landscape continues to evolve, the limits and restrictions we place today must be flexible enough to facilitate ethical data practices while not obstructing important advances.

Data ethics, as a field, is "an emerging branch of applied ethics which describes the value judgements and approaches we make when generating, analysing and disseminating data."[310] Data ethics can help reconcile the concerns that arise at the leading edge of this data landscape.[311] Per Google Director Rajen Sheth:

> [E]thical design principles can be used to help us build fairer machine learning models. Careful ethical analysis can help us understand which potential uses of vision technology are inappropriate, harmful, or intrusive. And ethical decision-making practices can help us reason better about challenging dilemmas and complex value tradeoffs—such as

[305] *See supra* Part III.B.1.

[306] *See* Hodge, *supra* note 201 ("It is a reality check on the limitations of any law, even a comprehensive one. Laws establish the minimum standard. Best practices are assumed to go beyond the minimum standard.").

[307] *See* Richards & King, *supra* note 133, at 429 (explaining that the law is slow to adapt to rapid change).

[308] *See id.* (stating that there can be a gap between "legal rules and the cutting-edge technologies that are shaping our societies").

[309] *See id.* at 396–97 ("Law will be an important part of Big Data Ethics, but so too must the establishment of ethical principles and best practices …."); IAPP, *supra* note 5, at 2; Devlin, *supra* note 259 ("Stronger legal frameworks, such as GDPR, and improvements in privacy technology can, at best, somewhat mitigate the risks. Ethics training and oversight for all business and IT personnel involved in the commissioning and implementation of big data analytics programs is not only necessary. It is the right thing to do.").

[310] U.K. DEP'T FOR DIG., CULTURE, MEDIA & SPORT, DATA ETHICS FRAMEWORK 3 (2018); Luciano Floridi & Mariarosaria Taddeo, *What Is Data Ethics?*, 374 PHIL. TRANSACTIONS ROYAL SOC'Y A 1, 1 (2016) (defining data ethics as "a new branch of ethics that studies and evaluates moral problems related to data …, algorithms … and corresponding practices … in order to formulate and support morally good solutions (e.g. right conducts or right values)").

[311] *See* Price & Cohen, *supra* note 89 (explaining that the rise of "big data" has increased risks to patient privacy).

> whether to prioritize transparency or privacy in an AI application where providing more of one may mean less of the other.[312]

At its heart, data ethics addresses the question: Just because something can be done, should it?[313]

This nascent field of ethics—defined as recently as 2016—has generated calls for the development of a comprehensive data ethics framework that can guide data decision-making.[314] Disparate parties have taken initial steps toward compiling ethical principles. Approaches have ranged from proposals as simple as adopting the "do no harm" ethos to the extraordinarily complex.[315] In 2018, Microsoft released a 151-page book, *The Future Computed: Artificial Intelligence and Its Role in Society*, that proposes ethical principles to guide the development of artificial intelligence.[316] The U.K. government has established a data ethics framework for data used by the public sector.[317] *Asilomar AI Principles* have offered twenty-three principles to guide artificial intelligence that have been signed on to by thought leaders ranging from Stephen Hawking to Elon Musk.[318]

[312] Rajen Sheth, *Steering the Right Course for AI*, GOOGLE CLOUD BLOG (Nov. 5, 2018), cloud.google.com/blog/products/ai-machine-learning/steering-the-right-course-for-ai.

[313] *See* Leetaru, *supra* note 48 ("In a world in which software can be built to do almost anything a programmer might imagine, the real question today is not what CAN we build, but rather what SHOULD we build?"); *see also* Mozilla, *supra* note 5 ("In a world where software is entwined with much of our lives, it is not enough to simply know what software can do. We must also know what software should and shouldn't do ….").

[314] *See* DIGITAL DECISIONS, *supra* note 6 ("A framework of principles is the first step in developing actionable solutions for the problem of biased automation, but it is far from the last."); IAPP, *supra* note 5, at 2; Floridi & Taddeo, *supra* note 310 ("[D]ata ethics should be developed from the start as a macroethics, that is, as an overall framework that avoids narrow, ad hoc approaches and addresses the ethical impact and implications of data science and its applications within a consistent, holistic and inclusive framework.").

[315] Lucy C. Erickson et al., *It's Time to Talk About Data Ethics*, ORACLE DATA SCI. BLOG (Mar. 26, 2018), https://www.datascience.com/blog/data-ethics-for-data-scientists ("One idea that has gained traction is the need for a 'Hippocratic Oath' for data scientists. Just as medical professionals pledge to 'do no harm,' individuals working with data should sign and abide by one or a set of pledges, manifestos, principles, or codes of conduct."); Gry Hasselbalch & Pernille Tranberg, *Data Ethics—The New Competitive Advantage*, TECHCRUNCH (Nov. 12, 2016, 7:00 PM), https://techcrunch.com/2016/11/12/data-ethics-the-new-competitive-advantage/; Tom Simonite, *Should Data Scientists Adhere to a Hippocratic Oath?*, WIRED (Feb. 8, 2018, 7:00 AM), https://www.wired.com/story/should-data-scientists-adhere-to-a-hippocratic-oath/ ("Microsoft released a 151-page book last month on the effects of artificial intelligence on society that argued 'it could make sense' to bind coders to a pledge like that taken by physicians to 'first do no harm.'").

[316] Brad Smith & Harry Shum, *The Future Computed: Artificial Intelligence and Its Role in Society*, OFFICIAL MICROSOFT BLOG (Jan. 17, 2018), https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/.

[317] U.K. DEP'T FOR DIG., CULTURE, MEDIA & SPORT, *supra* note 310.

[318] *See Asilomar AI Principles*, FUTURE LIFE INST., https://futureoflife.org/ai-principles/?cn-reloaded=1 (last visited Sept. 6, 2019).

Perhaps the most promising of these initial proposals is a collaboration between Bloomberg, BrightHive, and Data for Democracy to develop a code of ethics for data scientists from the ground up.[319] The Global Data Ethics Project is being developed as a community driven, crowd-sourced effort. The initiative has thus far produced a "living document" that establishes four values and ten principles.[320] At this inflection point, the time is right for a consensus ethical framework.

B. *The CRAFT Framework*

The ethical framework proposed below—the CRAFT framework—is the result of a comprehensive analysis of the existing data landscape, careful consideration of the strengths and weaknesses of existing ethical frameworks, and a robust review of the data ethics literature. The five principles of the CRAFT framework—Choice, Responsibility, Accountability, Fairness, and Transparency—can guide ethical data decision-making with an approach accessible to stakeholders across the data landscape.[321]

Unlike frameworks that address one aspect of data—algorithmic decision-making[322] or artificial intelligence[323]—this proposal is meant to address activities across the data lifecycle. The principles are meant to be specific and action-guiding, as opposed to aspirational; comprehensible to all data decision-makers, even those without formal ethics training; and adaptable enough to address the spectrum of emerging and evolving ethical concerns.

Choice is the ethical principle that grounds the CRAFT framework. Choice recognizes that individuals have ongoing interests in the uses of their data that must be taken into account. Choice is distinct from other generally proposed

[319] *Bloomberg, BrightHive, and Data for Democracy Launch Initiative to Develop Data Science Code of Ethics*, CISION PR NEWSWIRE (Sept. 25, 2017), https://www.prnewswire.com/news-releases/bloomberg-brighthive-and-data-for-democracy-launch-initiative-to-develop-data-science-code-of-ethics-300524958.html.

[320] *The Global Data Ethics Project*, DATA FOR DEMOCRACY, https://datapractices.org/community-principles-on-ethical-data-sharing/ (last visited Sept. 6, 2019).

[321] *See*; IAPP, *supra* note 5, at 8 ("In light of the myriad ethical issues raised by data analytics and AI, professionals working with organizations using big data should have a basic understanding of data ethics and tools for incorporating it into decision making."); Richards & King, *supra* note 306, at 430 ("Big Data Ethics needs to be part of the professional ethics of all big data professionals, whether they style themselves as data scientists or some other job description."); Romero, *supra* note 94 ("What *is* clear is that incidents like this one highlight the need for enhanced scrutiny and critical thinking from everyone involved—from app developers and researchers to everyday people agreeing to share their data.").

[322] DIGITAL DECISIONS, *supra* note 6.

[323] *Asilomar AI Principles*, *supra* note 318.

approaches to user decision-making including consent[324]—the approach taken in the Belmont Report—and control.[325] Requiring consent for the thousands of actions that give rise to data is too onerous to do meaningfully and meaningless if just another box to check.[326] At the other end of the spectrum, granting individuals control or ownership over their data can create significant roadblocks that hamper innovation while burdening individuals with the responsibility of data decision-making. Choice, as articulated in the CRAFT framework, is more involved than the mere opt-in or opt-out provisions required by FIPPs.

Rather, under CRAFT, individuals must be given a choice, and have their choices honored, when practices result from a transaction that: (1) could have meaningful consequences for the individual, and for which reasonable people could have differing preferences; and/or (2) are not within the bounds of an individual's reasonable expectations of the transaction. One such example is in the world of direct-to-consumer (DTC) genetic testing. There, individuals submit a sample of DNA for analysis and get results about their ancestry or health. What often happens, unbeknownst to consumers, is that DTC genetic testing companies de-identify samples,[327] and then share the de-identified samples with third parties without consumers' knowledge or consent.

In practice, the ability to take essentially unlimited action with genetic data turns on whether the data has been de-identified. Even identifiable data that falls outside the scope of narrow sectoral privacy laws can generally be shared without consent. Under a notice and consent approach, users can be provided notice in impenetrable legal language and be required to check a box before proceeding, requiring users to opt in to access the service. These "safeguards," such as they are, are insufficient.

Under the CRAFT framework, the ability to share data with third parties that are unrelated to the data processing without informing consumers does not turn on whether the data has been de-identified. Rather, because sharing with external third parties could have meaningful consequences to the consumer and would be outside the expected scope of the transaction, the sharing of either identifiable

[324] Romero, *supra* note 96 ("The basic concept of data privacy is built around the assumption that it's possible to consent to the way your data is being used. The Strava case makes it clear that such an assumption might be outdated, as it's fueled a belief held by many in the data privacy sphere that individuals can't consent to use cases of their data they don't yet know about.").

[325] *See* Richards & King, *supra* note 307, at 412, 421 (defining FIPs and noting that FIPs aim "to provide individuals control over their personal data so that they can weigh the benefits and costs at the time of collection, use, or disclosure").

[326] *See supra* Part IV.A.

[327] *See supra* Part III.C (discussing the limits of de-identification).

or de-identified data would require soliciting and honoring a consumer's choice. As captured by principles of Responsibility and Accountability, those directly transacting parties would have responsibilities for ensuring that user choice is honored across the data landscape; this would require taking into consideration whether user choice could appropriately be honored across the data landscape before making a decision to share data with third parties.

Data decision-makers at unrelated third parties would be expected to take these ethical principles under advisement at critical data decision-making junctures. In the context of less visible data collection mechanisms, like Internet cookies, implementation could look different. This may mean that browsers solicit user choice at one point and allow or prevent cookie tracking across the Internet in accordance with those preferences. Ultimately, data decision-makers confronted with developing the architecture of this evolving landscape should take the principle of honoring Choice at critical junctures seriously.

The second and third principles—Responsibility and Accountability—address the networked nature of data flows and the relationships among actors throughout the data landscape that were not foreseeable at the time the Belmont Report and FIPPs were developed. The second principle, Responsibility, recognizes that data decision-makers are responsible to individuals across the data landscape, even if they never directly interact. Should a data subject share data with a company, that subsequently shares that data with a third party in a manner permissibly within the bounds of the first transaction, the third party is nevertheless responsible to the data subject despite not directly transacting. In the DTC genetic testing example, if a consumer chooses to allow their genetic material to be shared with third-party researchers, the third-party researchers still owe responsibilities to the data subject, even if not in direct privity. The principle recognizes that the easy transmissibility of networked data often makes invisible the link between data subject and data decision-maker; decision-makers are nevertheless responsible to the data subjects.

The third principle, Accountability, holds data practitioners accountable for their acts and omissions and the downstream consequences that flow therefrom.[328] When nonconsensual data sharing permissibly falls within the scope of consumer expectations, such that consumers' choices need not be solicited, those who share data with third parties are nevertheless accountable for the reasonably foreseeable downstream data misuse that results from their

[328] IAPP, *supra* note 5, at 12 ("Accountability is the backbone of any data management program, including where data analytics—and therefore ethical issues—are involved."); Hodge, *supra* note 200.

decision to share. Those who profit from decisions to share data should bear some of the consequences rather than having the consequences fall solely on data subjects who were never consulted. In the example of DTC genetic testing, a DTC genetic company who, within the bounds of the transaction, chose a third-party sequencing company with inadequate data safeguards that subsequently had a preventable data breach would be accountable for the consequences even though the DTC company did not itself have the breach. In a networked data landscape, it is eminently foreseeable that decisions to share data could have far-reaching consequences; companies should take those into account and weigh them appropriately before deciding to share data.

Fairness, the fourth principle, includes both fairness to society and fairness to individuals.[329] In this sense, the notion of fairness is broader than the requirements of distributive justice set forth in the Belmont Report. It is also a mechanism for asking not just the question of "when is this legally permissible?" but also "should this be done?" Fairness to society requires that practitioners recognize and interrogate ways that biases become encoded into algorithms and machine learning, and work to prevent exacerbating biases as a result of these technologies.[330] Fairness to individuals grants individuals the right to fair processes and the fair transaction they reasonably thought they had entered into. Interactions between subject and practitioner convey to the subject the scope of the relationship and the types of data activities that will result; fairness requires that practitioners engage with subjects in ways that are not misleading. Sometimes the two notions of fairness intersect. Amazon's discriminatory hiring algorithm violated notions of societal fairness by systematically preferring men over women; it also violated individual fairness with respect to the specific women who would have been selected for positions but for the algorithmic decision-making. Fair processes would have had data decision-makers looking for bias before implementing the automated system and, upon finding evidence of bias, taking it down, as was done by Amazon. Fair processes could also be operationalized in other ways, including by granting individuals affected by automated decisions a right to have those decisions reviewed by a human, as is offered in the GDPR.[331]

[329] Hodge, *supra* note 201.

[330] *See* DIGITAL DECISIONS, *supra* note 6 ("Computerized decision-making … must be judged by its impact on real people, must operate fairly for all communities, and in particular must protect the interests of those that are disadvantaged or that have historically been the subject of discrimination."); Cathy O'Neil, *The Ethical Data Scientist*, SLATE (Feb. 4, 2016, 8:30 AM), https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html ("The ethical data scientist would strive to improve the world, not repeat it. That would mean deploying tools to explicitly construct fair processes.").

[331] *See* GDPR, *supra* note 184 2016 O.J. (L 119).

The final principle, Transparency, requires openness by practitioners about the paths that data can travel and the aims and goals of the data initiatives.[332] Principles of transparency have traditionally revolved around providing notice—requiring disclosure about collection, sharing, and use prior to collection. Entities discharge their notice obligations by disclosing as much information as possible, often using dense, legal language that is impenetrable to the average reader. Transparency under the CRAFT framework means operating under a different governing ethos—deliberately shining light on the logic of the decision-making process[333] and "giv[ing] users enough information to assess how much trust they should place in its digital decisions."[334] This does not mean, however, that every granular potential future use must be disclosed; in fact, quite the opposite. Because the Transparency principle recognizes that much about the path data will travel is unknown at the point of collection, uncertainty might be part of the disclosure. The operating principle, however, is that data practitioners must provide data subjects with the types of information a reasonable data subject would want to know in a manner calculated to achieve understanding.

C. Operationalizing Data Ethics

Given that the proposed CRAFT framework is nonbinding, and some have worried that companies only enact ethical reforms when they fear consequences to their bottom lines,[335] how can ethical frameworks such as the CRAFT framework facilitate a data landscape that warrants the full faith and confidence of the public?

Even nonbinding ethical frameworks can "affirmatively create a community with common values" and establish baseline expectations.[336] As baseline consumer expectations of ethical data practices become established, market forces could generate pressure that encourages companies to comply with ethical

[332] *See* IAPP, *supra* note 4, at 12 ("[T]ransparency builds trust in the system[] by providing a simple way for the user to understand what the system is doing and why."); Richards & King, *supra* note 307, at 396 ("Transparency has long been a cornerstone of civil society as it enables informed decision making by governments, institutions, and individuals alike…. Transparency can help prevent abuses of institutional power while also encouraging individuals to feel safe in sharing more relevant data to make better big data predictions for our society."); Hodge, *supra* note 201; Volchenboum, *supra* note 57 (arguing that if companies are going to ask users to share data, they must be transparent).

[333] *See* DIGITAL DECISIONS, *supra* note 6.

[334] *Id.*

[335] *See* Simonite, *supra* note 315.

[336] Metcalf, *supra* note 80 ("Given the multiplicity of purposes fulfilled by ethics codes/policies it is worth noting that enforcement may not be as important as it first appears. Many of the purposes of professional ethics codes are positive: codes can affirmatively create a community with common values.").

data practices. Companies that engage in ethical data practices, such as incorporating the CRAFT framework into their data decision-making, can market themselves as being CRAFT-compliant. This could have a market advantage comparable to companies that engage in environmental practices that earn them an Energy Star label. Indeed, some "[v]isionary companies are already positioning themselves within this movement and investments in companies with data ethics are on the rise."[337] The inverse may also be true; companies that develop reputations for having poor data practices may suffer in the marketplace.[338]

Companies that choose to implement an ethical framework, such as CRAFT, can employ additional ethical safeguards. For example, companies could create a Chief Ethics Officer position.[339] A Chief Ethics Officer should serve in a leadership position, independent from outside pressure, and be able to consult with data decision-makers across the organization to identify concerns and propose solutions to executive leadership.[340] Companies could rely on ethics review boards—internal or external—to identify and address ethics issues raised by particular data practices,[341] an approach modeled on Institutional Review Boards as required by the Common Rule.[342]

Companies could invest in training initiatives to ensure that data scientists are sufficiently well-versed in ethics to appreciate the ethical concerns.[343] Data scientists need not be experts at resolving every aspect of the ethical challenges of their work, but should be able to identify concerning data practices and begin important conversations.[344] Although training and pedagogical materials for data ethics are still under development,[345] the hope is that the next generation of data

[337] Hasselbalch & Tranberg, *supra* note 315.

[338] *See* IAPP, *supra* note 5, at 12 (explaining how unethical companies in Japan are publicly shamed).

[339] *See id.* at 13 (arguing that data ethics needs an internal leader comparable to chief privacy officers).

[340] *See id.* at 13–14 (explaining the role of data ethics leaders in an organization).

[341] *Id.* at 10.

[342] IAPP, *supra* note 5; Victoria Berkowtiz, *Common Courtesy: How the New Common Rule Strengthens Human Subject Protection*, 54 HOUS. L. REV. 923, 925, 938 (2017).

[343] IAPP, *supra* note 5, at 14 ("Data ethics should be assessed at each point in the data life cycle …. To that end, every employee should have basic data ethics training."); DJ Patil, *Data Science at the NIH and in Healthcare*, MEDIUM (Apr. 3, 2018), https://medium.com/@dpatil/data-science-at-the-nih-and-in-healthcare-d11f591c3312 ("[E]very training program on data, needs to have ethics and security integrated into the curriculum.").

[344] *See* IAPP, *supra* note 5, at 14; O'Neil, *supra* note 330 ("A data scientist doesn't have to be an expert on the social impact of algorithms; instead, she should see herself as a facilitator of ethical conversations and a translator of the resulting ethical decisions into formal code. In other words, she wouldn't make all the ethical choices herself, but rather raise the questions with a larger and hopefully receptive group.").

[345] *See* IAPP, *supra* note 5, at 8 ("Big data draws on the fields of physics, computer science, and applied mathematics, disciplines that 'have not been required to practically grapple with ethics requirements, [and

scientists will be more aware of the power that technology has to shape society and the responsibility that this entails.[346]

CONCLUSION

We have entered a world, unwittingly and unaware, in which many of our activities—online and offline—are monitored. Data about these interactions are collected, aggregated, shared, and subjected to complex data analytics. These analytic tools hold out tremendous potential for benefit: earlier disease diagnosis, solutions to public health emergencies, and better allocation of scarce health resources. But data analytics also raise troubling concerns—the size of the datasets being analyzed and the speed with which decisions are made give data analytics the power to amplify and exacerbate existing inequities. Data analytics produce outcomes that are impossible to query because of the opaque, black box nature of their processes. And individuals are given no meaningful way to opt out or engage in self-protection.

Although the law could offer protections—limiting unethical data practices and granting individuals rights in their data—existing laws fall woefully short. The United States has historically taken a sectoral approach to privacy law, granting privacy protections to extremely narrow types of data. None of these laws are able to address the concerns of networked data that are easily transmitted across boundaries. More comprehensive privacy laws recently enacted or proposed also fall short at protecting individuals from unethical data practices. And all privacy laws exempt de-identified data from protections altogether—a glaring exemption in an era of networked datasets in which re-identification is trivial. The result is that all of us are massively under-protected by existing law.

Given these limitations, some have turned to ethical frameworks to guide data decision-making. These frameworks—the Belmont Report and the Fair

therefore] they often lack access to pedagogical resources about research ethics that are widespread in other fields.'" (quoting Metcalf et al., *supra* note 79)); Natasha Singer, *On Campus, Computer Science Departments Find Blind Spot: Ethics*, N.Y. TIMES, Feb. 13, 2018, at B4 ("This semester, Harvard University and the Massachusetts Institute of Technology are jointly offering a new course on the ethics and regulation of artificial intelligence…. And at Stanford University, the academic heart of the industry, three professors and a research fellow are developing a computer science ethics course for next year."); Leetaru, *supra* note 48 ("[T]he initiative will award up to $3.5M to 'promising approaches to embedding ethics into undergraduate computer science education, empowering graduating engineers to drive a culture shift in the tech industry and build a healthier internet.'"); Metcalf et al., *supra* note 80 ("[T]he field of data science is finding that it does not have the ethics curricula or training materials developed for handling ethical challenges.").

[346] IAPP, *supra* note 5, at 2 ("Big data, new technologies, and new analytical approaches, if applied responsibly, have tremendous potential to be used for the public good.").

Information Practices Principles—were developed in the 1970s and are ill-equipped to address the concerns raised by networked data. Without meaningful ethical guidance, an industry with a governing ethos of "move fast and break things" will continue to break the most significant tenets of society—democracy, identity, and autonomy.

Recognizing that the data landscape is at an inflection point, this Article proposes the CRAFT framework—explicating ethical principles of Choice, Responsibility, Accountability, Fairness, and Transparency—to guide data decision-making and provide a foundation for subsequent legislation and comprehensive data governance.