Harvard Journal of Law & Technology
Volume 30, Number 1 Fall 2016

GET TO KNOW ME: PROTECTING PRIVACY AND AUTONOMY UNDER BIG DATA'S PENETRATING GAZE

*Sheri B. Pan**

TABLE OF CONTENTS

I. INTRODUCTION

Big data, the storage and analysis of large datasets, now affects everyday life.[1] It personalizes ads, calculates criminal sentences, and predicts criminal activity or, recast in a different light, constructs filter

* Judicial law clerk; Harvard Law School, J.D. 2016. The views expressed in this Note are solely my own and do not reflect the views of others. I thank Professor Urs Gasser for advising the paper that led to this Note, Kayla Haran for her comments as Article Editor, and the staff of the *Harvard Journal of Law & Technology* for their insight and diligence. This Note is dedicated to girls and women.

1. *See* Jonathan Stuart Ward & Adam Barker, *Undefined by Data: A Survey of Big Data Definitions*, ARXIV:1309.5821 (Sept. 2013), http://arxiv.org/pdf/1309.5821v1.pdf [https://perma.cc/768K-7VJG].

bubbles,[2] violates rights of procedural due process, and enables police departments to target communities on a discriminatory basis.[3] Both the benefits and dangers of the applications of big data have been widely discussed in popular discourse and legal literature.[4] But before big data can be used by companies and governments to provide services or make decisions, it must first derive inferences about the people within datasets. It compiles, analyzes, evaluates, and predicts a person's actions and attributes, all before the conclusions are used for a business or state purpose.

Current privacy discussions are predominantly concerned with how inferred information is used.[5] This Note, however, proposes that the process of analyzing data to infer information about people also threatens their privacy and autonomy interests. This Note proceeds in four parts: Part II summarizes current academic, legal, and industry conceptions of informational privacy and argues they have failed to consider the harm potentially posed by big data's capability of inferring new personal information; Part III considers the novel and unique characteristics of big data collection and analytics; Part IV discusses how big data threatens privacy and autonomy interests by making inferential conclusions about people's attributes and conduct, even if the conclusions are never used; and Part V proposes a framework to differentiate between data analysis that is innocuous and harmful. The framework states that a data-mining algorithm violates privacy and autonomy interests if: (1) it relies on an unexpected correlation between data points, (2) it infers personal information of a particularly sensitive nature, and (3) generating the inference breaches contextual integrity.

---

2. Filter bubbles result when websites personalize content, such as newsfeeds and search results, to reflect a person's tastes. As a result, the person obtains less exposure to viewpoints, ideas, or people that she disfavors. *See How to Burst the "Filter Bubble" that Protects Us from Opposing Views*, MIT TECH. REV. (Nov. 29, 2013), https://www.technologyreview.com/s/522111/how-to-burst-the-filter-bubble-that-protects-us-from-opposing-views/ [https://perma.cc/Q4S6-U9QQ].

3. *See, e.g.*, Maurice Chammah, *Policing the Future*, THE VERGE (Feb. 3, 2016), http://www.theverge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project [https://perma.cc/3SH4-CNBL] (police patrols); Eli Pariser, *Beware Online "Filter Bubbles,"* TED (Mar. 2011), https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles [https://perma.cc/FX4B-8P59] (filter bubbles); *infra* Part IV (procedural due process); *see also* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 809–10 n.11 (2014) (procedural due process).

4. *See generally, e.g.*, Gary D. Bass, *Big Data and Government Accountability: An Agenda for the Future*, 11 I/S: J.L. & POL'Y FOR INFO. SOC'Y 13 (2015); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework To Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41 (2013); *Big Data and the Future of Privacy*, EPIC.ORG, https://epic.org/privacy/big-data/ [https://perma.cc/FX4E-JWB5].

5. *See infra* Part II.

II. CURRENT CONCEPTIONS OF PERSONAL INFORMATION

Privacy has traditionally been difficult to define and regulate. Despite disagreement over how to best treat the issue, privacy theories, privacy law, and privacy policies share a characteristic in common: conceptualizing personal information as static pieces of knowledge about someone. Part II makes this observation by examining theories of privacy, privacy laws, and privacy policies.

A. Privacy Theories

A fundamental theory of privacy defines privacy as the control over personal information. In his seminal book on privacy, privacy scholar Alan Westin articulates the control theory as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."[6] Legal scholar Arthur Miller writes that privacy is "the individual's ability to control the circulation of information relating to him."[7] In other words, the privacy-as-control perspective concludes that a person maintains privacy when she can decide how her information is collected, shared, used, retained, or otherwise manipulated.

Before big data, maintaining control over the data one shared with others necessarily meant controlling one's personal information. If a viewer voluntarily gave Netflix her ratings of certain movies and decided how Netflix could share, use, and retain the ratings, she maintained control over the information. But how does the control theory evaluate privacy where some personal data, voluntarily collected from the person, can be analyzed to infer other information never disclosed by the person? Imagine a situation where the viewer similarly discloses her movie ratings, but those movie ratings when analyzed by an algorithm can predict that the viewer is likely homosexual. As in the traditional paradigm, she voluntarily shared her movie ratings. But has she nonetheless maintained control over her personal information and has Netflix respected her privacy interests, given that she never agreed to reveal her sexual orientation? The privacy as control theory provides no clear answer, because it views control over personal information as control over the information a person has shared with others.

In recent years, Helen Nissenbaum's contextual integrity theory of privacy has become highly influential.[8] The contextual integrity framework defines privacy based on the norms that govern infor-

6. ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967).

7. ARTHUR MILLER, THE ASSAULT ON PRIVACY 25 (1971).

8. *See generally* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 101 (2004).

mation in different contexts. Privacy is preserved where an entity collects and shares information about a person in a way that comports with the expectations of that context.[9] In the context of healthcare, it is appropriate for a doctor to solicit a list of current medications from a patient. In an employment context, however, collecting the same information violates the right to privacy. The contextual integrity theory thus recognizes that the context of information collection and sharing plays an important role in whether the collection or sharing is appropriate, and is more nuanced than the control theory. Contextual integrity, however, still does not explicitly address situations where the information at issue changes. What if the doctor's office obtained a patient's list of medications, but then used the list to infer the likely race of the patient? Collecting the medication information is consistent with the norms governing a healthcare context, but it is nonetheless unclear whether generating the additional inference violates privacy rights. The theory does not readily provide an answer, because the context at issue, and thus the norms at issue, has remained the same. It is primarily concerned with differences in norms between contexts, rather than the generation of new information within the same context. Nowhere is this more clear than Nissenbaum's treatment of data mining as a case study. The discussion of data mining does allude to its ability to learn information about people. In examining the privacy harms posed by data mining, however, Nissenbaum contemplates only two types of situations — "the grocer who bombards shoppers with questions about other lifestyle choices" and "[t]he grocer who provides information about grocery purchases to [third parties]" — as ones that may pose privacy risks.[10]

B. Privacy Laws

In the United States, there is no overarching information privacy statute or common law regime,[11] but the Federal Trade Commission ("FTC") issues policy statements, initiates investigations, and reaches settlements to police privacy violations by companies.[12] The reports it publishes tend to treat information as static, although this is beginning to change in more recent reports on big data. In the 2012 report on Protecting Consumer Privacy in an Era of Rapid Change, the FTC presents data security, reasonable collection limits, sound retention practices, and data accuracy as the key principles to safeguarding pri-

9. *See id.* at 120–21.

10. *See id.* at 103, 134–35.

11. James P. Nehf, *Recognizing the Societal Value in Information Privacy*, 78 WASH. L. REV. 1, 58 (2003).

12. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 620–27 (2014).

vacy.[13] It recommends that companies provide consumer choice at the time of collection and use, limit sharing data with third parties, obtain affirmative consent before collecting sensitive information, and develop the Do Not Track mechanism.[14] By positing collection, use, sharing, and retention as the key bulwarks of privacy, the FTC assumes that controlling the original piece of information collected from a person is enough to protect privacy interests. A focus on collection, use, sharing, and retention does not contemplate that collected data can reveal additional inferences about a person or how companies should treat those inferences. Recent reports by the FTC dedicated to big data and data brokers do acknowledge that existing data can be used to learn additional information about a person.[15] The FTC may thus be starting to conceptualize information as more than static pieces of knowledge.

C. Privacy Policies

Privacy policies by companies generally focus on what is done to each piece of collected information and do not address the privacy implications of using collected data to generate new inferences. Google's Privacy Policy is organized into sections on information collection, use, transparency and choice, access, sharing, and security, and does not have a section dedicated to data analysis.[16] It does state that Google uses "automated systems [to] analyze . . . content" in order to customize search results, tailor advertising, and detect spam, and that Google will not make sensitive inferences such as race and religion.[17] Those statements, however, are the extent to which the privacy policy discloses the company's capacity to learn new information about users of its products. Facebook's privacy policy is

---

13. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 23 (2012), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf (last visited Dec. 16, 2016).

14. *See id.* at 35–60.

15. *See* FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? ii, 3 (2016) ("The life cycle of big data can be divided into four phases: (1) collection; (2) compilation and consolidation; (3) data mining and analytics; and (4) use."), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf [https://perma.cc/28WX-U7LC]; FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY vi, 47 (2014) ("Data Brokers Combine and Analyze Data About Consumers to Make Inferences About Them, Including Potentially Sensitive Inferences . . ."), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/547Q-LTHM].

16. *See Privacy Policy*, GOOGLE (Aug. 29, 2016), https://www.google.com/policies/privacy/ [https://perma.cc/4DP4-X5AA].

17. *Id.*

similar. It is divided into sections on information collection, use, sharing, control, and disclosure to the government.[18] It states that Facebook "analyze[s] the information [it has]" to improve products and services, conduct audits and troubleshooting activities, and improve advertising, but does not otherwise discuss the ways in which the company infers new information about people.[19] The same is true of guides to drafting privacy policies. The Better Business Bureau advises businesses to disclose their practices regarding information collection, use, user control, and security, and the California Attorney General recommends sections on data collection, online tracking, use, sharing, individual choice and access, and security.[20] Neither guide discusses whether and how companies should explain the ways they analyze collected information to make additional conclusions about people's attributes or conduct.

Current privacy theories, laws, and policies do not directly deal with how to evaluate privacy where new knowledge is inferred, because they were designed to tackle data collection and analysis before the advent of big data. How big data has changed information gathering and analysis is the subject of Part III.

III. CHARACTERISTICS OF BIG DATA

Companies have collected and analyzed personal information for decades.[21] Big data most obviously differs from previous methods of information processing in that it involves larger volumes of data and more powerful analysis.[22] It also differs, however, in various qualitative ways. This Part discusses four main areas in which big data treats personal information in a novel manner: first, collection occurs constantly and imperceptibly; second, what an organization knows about a person can grow as data analytics generate new inferences about the person; third, data analytics can infer sensitive details about a person using innocuous personal data; and fourth, the inferences generated about a person can be unpredictable, because the correlations that big data discovers between data are often inexplicable.

---

18. *See Data Policy*, FACEBOOK (Jan. 30, 2015), https://www.facebook.com/policy.php [https://perma.cc/R8ZD-PECN].

19. *Id.*

20. CAL. DEP'T OF JUSTICE, MAKING YOUR PRIVACY PRACTICES PUBLIC 10–14 (2014), https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making_your_privacy_practices_public.pdf [https://perma.cc/K5HB-XFFF]; *Tips on Establishing a Privacy Policy*, BETTER BUS. BUREAU, https://www.bbb.org/dallas/for-businesses/bbb-sample-privacy-policy1/tips-on-establishing-a-privacy-policy/ [https://perma.cc/QU73-HX2Y].

21. *See* Thomas H. Davenport, *Analytics 3.0*, HARV. BUS. REV. (Dec. 2013), https://hbr.org/2013/12/analytics-30 (last visited Dec. 15, 2016).

22. *See* Ward & Barker, *supra* note 1.

A. Data Collection Is Constant and Imperceptible

Big data operates by analyzing large datasets, so it relies on obtaining large volumes of information.[23] In developed economies today, data collection occurs constantly. This is in large part due to the prevalence of objects that can sense, store, and transfer information. Smart meters, for example, can collect and transmit a home's energy usage throughout the day, hourly or even more frequently.[24] Fitbits track the movement of their wearers during the day and night, from the metrics of a morning jog to bedtime sleep patterns. These objects do not turn off their recordings, but collect streams of information.

As more collection becomes constant, it is also increasingly imperceptible. Devices automatically take measurements without human intervention. They do not ask the data subject for consent or provide notice every time they record a reading. Because data collection occurs in the background, it easily goes unnoticed. This is especially true of wearables like the Fitbit. After initially putting one on, wearers need not think about the device for it to track, label, and store their movements. Fitbit emphasizes its seamless integration into people's lives, noting on its website that the device "automatically" records physical activities, monitors sleep, and syncs data online.[25]

Fitbit, like most technology companies, provides a privacy policy that informs consumers what, when, and how a service or product collects personal information. However, most people do not read legal disclaimers or terms of use.[26] Even assuming that consumers do read, understand, and agree to legal notices, knowing that personal data will be collected in the future does not equal fully realizing the extent of collection. Some Fitbit wearers, for instance, have remarked that they forget to remove the tracker during sex or masturbation, leading to erroneous recordings that they had gone on a jog.[27] People can learn at some point when collection will occur, but nonetheless forget that collection is occurring when they use the product.

It is probably even more difficult for people to be aware of how much they are being tracked in spaces equipped with sensor networks.

---

23. *Id.*

24. FERC, ASSESSMENT OF DEMAND RESPONSE & ADVANCED METERING 5 (2008), http://www.ferc.gov/legal/staff-reports/12-08-demand-response.pdf [https://perma.cc/8PHB-5ZWL].

25. *See, e.g.*, *Fitbit Alta*, FITBIT, https://www.fitbit.com/shop/alta [https://perma.cc/8DXJ-ZCHC].

26. *See* Rainer Böhme & Stefan Köpsell, *Trained To Accept? A Field Experiment on Consent Dialogs*, *in* PROC. OF THE ACM CONF. ON HUM. FACTORS IN COMPUTING SYS., Apr. 2010, https://www.wi1.uni-muenster.de/security/publications/BK2010_Trained_To_Accept_CHI.pdf [https://perma.cc/PJ3P-R68F].

27. *See* Fitbater, *Forgot To Take It Off...*, REDDIT (Nov. 7, 2013), https://www.reddit.com/r/AdviceAnimals/comments/1q4o37/forgot_to_take_it_off/ [https://perma.cc/2U8N-JDRF].

Sensor networks contain multiple sensor devices that communicate with each other to monitor the environment.[28] Songdo, Korea is a smart city that contains sensor objects on vehicles to manage congestion and in streets to track foot traffic.[29] Smart homes can read the temperature, light level, and movement within each room. [30] Sensor networks record numerous types of information about inhabitants as they live their daily lives, and thus make it difficult, if not impossible, to always remain aware of the personal information collected at any given time.[31]

B. New Insights Are Generated

At the most basic level, big data departs from traditional data collection and analysis in that it generates new insights from data. Before big data, the universe of information that an entity had about a person was limited to what it had gathered. To construct a more complete profile, the entity had to obtain additional facts from the person herself or a third party. Although researchers before big data did analyze existing datasets to uncover patterns, datasets were small and analysis took weeks to months.[32] With the development of big data, organizations gained the capacity to truly develop their knowledge of an individual. Big data was made possible by several technological advancements: greater memory, greater storage capacity, and more powerful analytics technologies (such as database sharding, NoSQL, MapReduce, Yarn, and Hadoop).[33] Big data is what infamously enabled Target to predict whether a customer was pregnant and in what trimester, without directly soliciting that information and based only on purchase history.[34] What an entity learns about a person also continues to expand if new data is added to existing databases, because

---

28. *See Wireless Sensor Network (WSN)*, TECHOPEDIA, https://www.techopedia.com/definition/25651/wireless-sensor-network-wsn [https://perma.cc/9T4X-DZFR].

29. Gerhard P. Hancke et al., *The Role of Advanced Sensing in Smart Cities*, 13 SENSORS 398, 416 (2013).

30. *See* Martin LaMonica, *Will Smart Home Technology Systems Make Consumers More Energy Efficient?*, THE GUARDIAN (Jan. 22, 2014), https://www.theguardian.com/sustainable-business/smart-home-technology-energy-nest-automation [https://perma.cc/FS3L-GW4P].

31. A smart home can generate 1 GB of data per week. Stacy Higginbotham, *How Much Data Can One Smart Home Generate? About 1 GB a Week.*, GIGAOM (Jul. 29, 2014), https://gigaom.com/2014/07/29/how-much-data-to-a-smart-home-generate-about-a-1-gb-a-week/ [https://perma.cc/W4FE-R2XE].

32. *See* Davenport, *supra* note 21.

33. *See* Ward & Barker, *supra* note 1; *Database Sharding*, AGILDATA, http://www.agildata.com/database-sharding/ [https://perma.cc/CN5M-QHF3].

34. *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (last visited Dec. 15, 2016).

algorithms can find new patterns between data points.[35] In that sense, big data has the ability to get to know someone over time.[36]

C. Inferred Information Is Often Sensitive

Big data is capable of using innocuous data about a person to make inferences of a sensitive nature.[37] This Note posits that big data derives an inference that is sensitive in two circumstances: where an algorithm makes a qualitative conclusion about someone using quantitative data or makes an inference related to a person's attribute or demographic characteristic using data on her behavior.[38] Figure 1 depicts the sensitivity of information on a spectrum, from less sensitive to more sensitive. Quantitative data points and behavior-related information are less sensitive, while qualitative conclusions and attribute-related information are more sensitive.

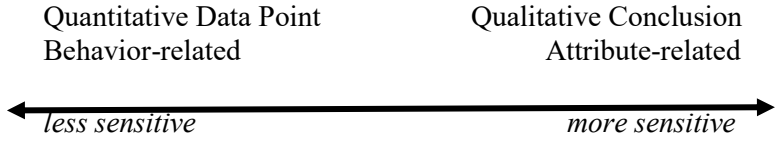

Figure 1: Sensitivity Spectrum

Big data often involves the analysis of a set of individual measurements to draw qualitative conclusions. Quantitative data is generally less sensitive. It is information that pertains to a single point in time and is not a conclusion about who the data subject is overall. For example, a recording that a person's home was using 3.2 kilowatts at 6:15 PM on a particular day does not reveal much about the person. A

35. *See* MAX BRAMER, PRINCIPLES OF DATA MINING 5–6 (2d ed. 2013).

36. Big data, of course, does not always infer information that is accurate. Google notoriously failed to forecast flu prevalence using only search queries. *See* David Lazer & Ryan Kennedy, *What We Can Learn from the Epic Failure of Google Flu Trends*, WIRED (Oct. 1, 2015), http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends [https://perma.cc/S328-FQTD]. However, many algorithms are highly accurate. Machines are better than humans at predicting student grades and parole violation rates, for instance. *See* LUKE DORMEHL, THE FORMULA: HOW ALGORITHMS SOLVE ALL OUR PROBLEMS... AND CREATE MORE 119–21, 208, 212 (2014).

37. This Part avoids defining sensitivity using subject matter categories such as health, sexuality, or finances because academics differ on which categories constitute sensitive information. *See, e.g.*, Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1138, 1150–57 (2015).

38. Conclusions generated by big data may always be more sensitive, in a sense, because they were not voluntarily shared by the person. If a person agrees to disclose a piece of information, she was presumably comfortable sharing it. By contrast, a person may not have wanted to share a piece of information inferred about her. This observation, however, is ultimately unhelpful to defining when big data generates conclusions of greater sensitivity, because all algorithms aim to infer new information.

qualitative judgment about a person, by contrast, does disclose substantive personal information. A home's energy usage can predict the resident's daily routine, her level of income, and whether she grows marijuana inside her home.[39] One outdoor running route, recorded by GPS, is low sensitivity information, but analyzing a history of jogging routes will likely uncover the runner's home address, high sensitivitity information.[40] Quantitative measurements are not always innocuous: a bank account balance is sensitive, despite being an isolated and numerical data point. However, quantitative data in contexts like wearables, sensor networks, online browsing, and shopping are by and large less sensitive than the qualitative conclusions they can help generate.

Big data also analyzes information on a person's behavior to infer information about the person's attributes or demographic profile. Information on a person's actions may or may not be sensitive, depending on the conduct, but individual attributes and demographics are always sensitive as a general matter. This type of data includes a person's personality traits, degree of intelligence, employee value, race, gender, political identification, and the like. Some of these categories are sensitive because they are protected classes under the Constitution.[41] Other categories are sensitive because they speak to who a person is — her disposition, aptitude, background, outlook — and thus connect closely to personhood. For example, researchers analyzed the "likes" of Facebook users and discovered that liking the Facebook page "MAC Cosmetics," a makeup brand, is a strong predictor for homosexuality.[42] The act of liking a page is not very sensitive, but sexual orientation is highly sensitive information.

D. Discovered Correlations Are Unexpected

Big data can and does find relationships between data points that are surprising. For instance, data analysis reveals that the total revenue generated by arcades closely tracks the number of computer science doctorates awarded in the United States over the last ten years.[43] Data analysis finds surprising patterns because it uses inductive, rather than deductive, reasoning. Deductive reasoning uses one or more

---

39. *See, e.g.*, Kyllo v. United States, 533 U.S. 27 (2001). The police used thermal imaging on a house to determine it was emitting an unusually high amount of heat radiation and to obtain a warrant to search the property for marijuana plants. *Id.* at 27.

40. *Cf. Frequently Asked Questions*, STRAVA, https://www.strava.com/how-it-works [https://perma.cc/NYW8-F93Q] ("You can also create a privacy zone perimeter around any address like your home, office, or any place you tend to start activities from that you'd like to keep private.").

41. *See* United States v. Virginia, 518 U.S. 515, 533–54 (1996).

42. *See* Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PNAS 5802, 5804 (2013).

43. Tyler Vigen, *Spurious Correlations*, http://tylervigen.com/spurious-correlations [https://perma.cc/MB72-A5GG].

premises to reach a conclusion that must logically follow if the premises are true.[44] Inductive reasoning uses specific observations to infer a general conclusion that may or may not be true.[45] For example, classification algorithms analyze which values have most often appeared together in past instances in order to formulate a rule for when each value generally occurs and thereby predict future instances.[46]

A correlation, however, does not necessarily signal a cause and effect relationship. Correlation does not imply causation. This is one reason why big data discovers unexpected patterns between data points. In some cases, a third factor is the cause of a correlation: for example, an employment data analysis firm found that people who use manually installed browsers like Firefox rather than pre-installed browsers like Safari tend to perform better at work and stay longer at their jobs,[47] and a researcher speculated that choice of browser and work performance are both functions of a third data point, inclination to take initiative.[48] Some correlations, however, have no discernible explanation and yet are very strong. The study on Facebook likes found that one of the most accurate predictors of whether a user's race was white or black was whether she liked the Facebook page "Halloween."[49] It is unclear why.[50] Because these correlations have no discernible explanation, it is unlikely they would be detected without big data.

As storage capacity grows and data analytics becomes more powerful, big data will generate more unexpected inferences. Computers can explore datasets that are hundreds of terabytes in size, which humans cannot. Furthermore, some algorithms are designed so that it is impossible to anticipate what correlations they will find. Data analysis traditionally relied on supervised learning, which required a dataset labeled with values, for example the names and grades of students.[51] The goal of an algorithm was to predict the labeled values for unseen cases.[52] With advances in unsupervised learning, algorithms can now discover patterns in unlabeled datasets.[53] In 2012, Google's

---

44. *See Deductive and Inductive Arguments*, INTERNET ENCYCLOPEDIA OF PHIL., http://www.iep.utm.edu/ded-ind/ [https://perma.cc/6MG4-KLTW].

45. *See id.*

46. *See* BRAMER, *supra* note 35, at 5–6.

47. Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees*, THE ATLANTIC (Mar. 16, 2015), http://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ [https://perma.cc/6K58-9Q34].

48. *See id.*

49. Kosinski et al., *supra* note 42, at 5805.

50. Of course, the study discovered some correlations that were more intuitive. Liking "NOH8 Campaign," an organization that advocates for LGBTQ rights, is associated with being homosexual. *Id.* at 5804.

51. *See* BRAMER, *supra* note 35, at 4–5.

52. *Id.*

53. *See id.* at 5.

neural network analyzed 35,000 images from YouTube videos and identified the faces of cats — independent of direction from researchers to find cat videos.[54] Other unsupervised learning projects have succeeded in determining whether a social media image is emotionally positive, neutral, or negative,[55] and the boundaries of a person's separate social circles.[56] This and other characteristics of big data pose novel harms to people's privacy and autonomy. That discussion is the focus of Part IV.

IV. PRIVACY AND AUTONOMY HARMS FROM BIG DATA

Both popular and academic discourse have tackled the privacy and autonomy harms posed by big data. These discussions mainly focus on harms that result from how companies and the government use the conclusions made about people to provide services or make decisions. For instance, legal scholars have suggested developing "procedural data due process" to oversee how algorithms arrive at predictions.[57] The White House Council of Advisors on Science and Technology recommended regulating big data and privacy via limitations on how inferences are used, and not how data is analyzed.[58] This Part begins by summarizing the use harms that are the focus of current discourse on big data. It then delves into non-use harms and the striking challenges they present to privacy and autonomy.

A. Use Harms

The way that companies and government entities use the inferences produced by big data harms privacy and autonomy in three ways: over- and inaccurate personalization, violation of due process, and discrimination. Personalization is the tailoring of services to predictions of a user's individual preferences. It is displaying targeted

---

54. Quoc V. Le et al., *Building High-level Features Using Large Scale Unsupervised Learning* (2012), http://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf [https://perma.cc/HAD3-CJZM].

55. *See* Yilin Wang et al., *Unsupervised Sentiment Analysis for Social Media Images*, http://yilinwang.org/papers/Paper158_UESA.pdf [https://perma.cc/847S-HPDE].

56. *See* Julian McAuley & Jure Leskovec, *Learning To Discover Social Circles in Ego Networks* 1–2, https://cs.stanford.edu/people/jure/pubs/circles-nips12.pdf [https://perma.cc/T3RT-5UUN].

57. Crawford & Schultz, *supra* note 4, at 125–28.

58. PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY, BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE xiii (May 2014), https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf [https://perma.cc/6BP6-S8H8] ("Policy attention should focus more on the actual uses of big data and less on its collection and analysis. . . . [A] priori limitations on . . . analysis (absent identifiable actual uses of the data or products of analysis) are unlikely to yield effective strategies for improving privacy.").

ads, recommending movies of interest, and curating social media feeds. Over-personalization, however, can restrict autonomy by encouraging people to act on their "impulsive, present selves" rather than "future, aspirational selves."[59] During the Egyptian Revolution of 2011, Google responded to a search for "Egypt" by displaying vacation guides to some users and news of the Revolution to other users.[60] Returning travel tips gave people what they wanted in the short term, while returning information on the revolution educated people to their long-term benefit. The divergent universes of information provided to different people are termed filter bubbles.[61] Filter bubbles may have the effect of eroding civic engagement and increasing political polarization. Conversely, personalization that is not accurate enough can violate autonomy interests by nudging people to take actions they otherwise would not. LinkedIn populates a "Jobs You May Be Interested In" list for each user.[62] If it displays job openings that are unsuitable to or unwanted by a person, people may apply for and accept positions that they otherwise would not consider.

The use of big data can also harm privacy and autonomy by enabling companies and government entities to make decisions about people without providing procedural due process. Procedural due process requires the government to provide certain procedures before depriving people of life, liberty, or property.[63] Prison parole boards are beginning to rely on data analysis to assign recidivism risk scores to inmates eligible for parole and decide whom to release.[64] Insurance companies use big data to assess each applicant's risk in order to calculate premium rates.[65] Leaning on big data to help make decisions affecting people's lives means that human decision makers play a lesser role and cases are not evaluated on an individualized basis that accounts for unique circumstances. This especially threatens autonomy and privacy interests where the decision affects a person's liberty. Furthermore, because algorithms, what factors they evaluate, and how they arrive at their inferences are not disclosed,[66] neither the govern-

59. Pariser, *supra* note 3.

60. *Id.*; *see also* Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 COLO. TECH. L.J. 203, 214–15 (2015) (examining how in 2015 Facebook disproportionately featured Ice Bucket Challenge videos on newsfeeds while Twitter prioritized commentary on Ferguson protests).

61. Pariser, *supra* note 3.

62. *Jobs You May Be Interested In*, LINKEDIN, https://www.linkedin.com/help/linkedin/answer/11783/jobs-you-may-be-interested-in-overview [https://perma.cc/DH2U-QYXW].

63. U.S. CONST. amend. V.

64. *Prison Breakthrough*, ECONOMIST (Apr. 19, 2014), http://www.economist.com/news/united-states/21601009-big-data-can-help-states-decide-whom-release-prison-prison-breakthrough [https://perma.cc/LN6W-95GW].

65. *See FICO Insurance Scores*, FICO, http://www.fico.com/en/products/fico-insurance-scores [https://perma.cc/W7LP-9P53].

66. *See generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2016).

ment nor the public can exercise oversight over whether big data's conclusions are legal or fair.

Big data also enables organizations to discriminate against people based on race or other demographics without doing so openly. Until recently, Facebook's "ethnic affinity marketing solution" allowed an organization to place ads for housing that targeted some users but excluded those who had an "affinity" of races of African-American, Asian-American, and Hispanic.[67] The tool did not assign ethnic affinity based on a user's self-reporting but based on interests and activities that were proxies for race.[68] While Facebook has suspended ethnic targeting for ads offering housing, employment, and credit, it is still available for other types of ads. Beyond advertising, there is the concern that entities like landlords, insurance companies, and businesses will charge higher prices based on gender, ethnicity, class, and more.[69] Decisions that disparately impact different groups also inflict harms of a dignitary nature. Computer scientist Latanya Sweeney found that Google is twenty-five percent more likely to show an ad related to arrest records in response to a search for a common African-American name compared to a common Caucasian name.[70] Associating African-Americans with crime and incarceration contributes to embarrassment, stigmatization, and stereotyping of black communities.[71]

B. Non-Use Harms

Less discussed in discourse on big data is how big data harms privacy and autonomy interests even when the inferences it generates about people are not used to provide services or make decisions. This Note proposes that big data causes four main types of non-use harms: big data enables organizations to learn information about people that they would not have disclosed, restricts autonomy by judging people's conduct and character, impedes the possibility of acting anonymously,

---

67. Sapna Maheshwari & Mike Isaac, *Facebook Will Stop Some Ads from Targeting Users by Race*, N.Y. TIMES (Nov. 11, 2016) http://www.nytimes.com/2016/11/12/business/media/facebook-will-stop-some-ads-from-targeting-users-by-race.html (last visited Dec. 15, 2016).

68. *Id.*

69. *See, e.g.*, Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARV. BUS. REV. (Jan. 29, 2014), https://hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination/ (last visited Dec. 15, 2016).

70. Latanya Sweeney, *Discrimination in Online Ad Delivery*, SSRN (Jan. 28, 2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240 [https://perma.cc/SJ5G-RFJF].

71. THE SENTENCING PROJECT, RACE AND PUNISHMENT: RACIAL PERCEPTIONS OF CRIME AND SUPPORT FOR PUNITIVE POLICIES 13–26 (2014), http://www.sentencingproject.org/wp-content/uploads/2015/11/Race-and-Punishment.pdf [https://perma.cc/5ELE-Z9BU].

and undermines the tenet that each person is an individual who possesses agency.

1. Learning Private Information

Most obviously, big data threatens privacy by inferring information about a person that she neither intended nor wanted to reveal.[72] As discussed in Part III, connected devices record personal data constantly and imperceptibly, and it is difficult for a person to fully realize what personal information is being collected for analysis. Furthermore, because algorithms can find unexpected patterns in data, the data subject cannot predict what personal information big data will infer about her. As a result, people unknowingly share information about themselves they did not intend to disclose. For example, a person may not publicly like the Facebook page "Sephora" if they know it is correlated with low intelligence.[73] Big data's potential for inferring undisclosed information about someone is especially harmful to privacy interests when the undisclosed information is more sensitive than the data collected. It is less likely that a person intended to share a sensitive versus insignificant fact.

2. Limiting Autonomy

Big data restricts autonomy by generating conclusions about people's attributes and behavior, thereby making judgments about them. It resembles a panopticon. Jeremy Bentham conceptualized the panopticon as a model prison, comprised of a central watchtower that is encircled by prison cells.[74] The guard in the watchtower can see into the cells, but the prisoners cannot see into the watchtower.[75] For Michel Foucault, this arrangement creates the automatic functioning of power because the prisoners obey even absent force — they know they are visible to the watchtower, so they voluntarily follow prison rules.[76] The watchtower exerts an external gaze on the prisoners, but the prisoners internalize this gaze.[77] Foucault believed that this design could be incorporated into other contexts to control people, in schools to keep students quiet or in offices to maintain employee focus.[78]

---

72. The inference must be accurate, however, for it to pose a threat. This is increasingly common as big data advances. *See supra* note 36.

73. Kosinski et al., *supra* note 42, at 5804.

74. *See* JEREMY BENTHAM, THE PANOPTICON WRITINGS (1995); MICHEL FOUCAULT, DISCIPLINE AND PUNISH 200 (Alan Sheridan trans., Vintage Books 2d ed. 1995) (1977).

75. *See* FOUCAULT, *supra* note 74, at 200–02.

76. *Id.* at 200–01.

77. *Id.* at 202–03.

78. *Id.* at 202.

The panopticon exerts control over its subjects not only through surveillance but also the ability to make judgments about them. Foucault noted that the panopticon allows the guard in the watchtower to not only monitor people but also evaluate their aptitude, character, and behavior.[79] The experience of being evaluated effects changes in the subject herself, in self-perception and conduct. Frantz Fanon built upon this idea in his exploration of the white gaze.[80] In his homeland of Martinique where he was of the native race, he viewed himself through his own eyes.[81] Upon moving to France and entering the white world, however, he discovered that white men saw him as different, as an Other,[82] and deemed him as inferior, dangerous, stupid, uncouth, and ugly.[83] The realization that he was labeled as different changed his own self-opinion; he no longer felt like a "man among men" but viewed himself as others regarded him.[84] The white gaze also altered how he and other blacks behaved.[85] He observed some blacks whitening their skin, marrying white spouses, and speaking more like whites.[86] Thus, awareness of an external gaze that not only watches but also makes judgments impinges upon autonomy.

Big data exerts an external gaze on people in judging their characteristics and conduct. It predicts and labels people's tastes, demographic information, future behavior, and more. Companies can classify people in groups as specific as 34 year-old men who are in the $100,000 – $125,000 income bracket, newly engaged for 6 months, and likely to engage in conservative politics.[87] Awareness that big data is watching and scrutinizing decisionmaking may restrict people's inclination to exercise free choice. A person may watch fewer gore films if she suspects Netflix will classify her as violent or avoid searching for an embarrassing medical condition if she believes Google is tracking her search queries. As physical objects are increasingly networked, big data also intrudes on offline behavior. If a married person thinks that an algorithm can infer whether she is having an affair based on her regular nighttime visits to a motel, she may well begin varying the locations of her rendezvous.

Making predictions about people's future selves also constrains willingness to experiment. All people discover and reconstitute them-

79. *Id.* at 203.
80. FRANTZ FANON, BLACK SKIN, WHITE MASKS 89–119 (Richard Philcox trans., 1952).
81. *Id.* at 89–90.
82. *See, e.g.*, *id.*
83. *See id.* at 91–109.
84. *See id.*
85. *Id.* at 1–5, 91, 92, 96.
86. *See id.* at 1–5, 91, 96.
87. *See Ads Manager*, FACEBOOK, https://www.facebook.com/ads/manager (link only visible when logged in) (enabling the user to create an ad that targets an audience based on location, age, gender, languages, demographics, interests, and behaviors).

selves throughout life,[88] but some confront particularly fundamental and complex questions about who they are — those who may be transgender, queer, or multiracial are but a few examples. To develop identity, people need the room to try on masks for size, practice new roles, and revert back to old versions of themselves, free from external judgment.[89] An algorithm may have made a conclusion about a person based on data collected during a period of self-exploration, and knowledge of those conclusions may chill her willingness to experiment in the future. The freedom to develop one's individuality is particularly important in a democracy, where society functions on a diversity of viewpoints and independent thought.[90]

The persistent threat of scrutiny also obstructs the emotional release that occurs only when privacy exists. People play social roles in public, but cannot perpetually sustain those roles without respite to recharge in time alone or with close friends.[91] Big data intrudes on that solitude because it continuously collects and analyzes people's online and offline habits. Some racial minorities feel pressure to code switch when interacting with white people, or speak with the linguistic style of white communities, in order to fit in.[92] Devices that record conversations, such as digital assistants like Amazon's Alexa, may reduce the proclivity to switch to a more comfortable but less mainstream parlance even when at home. Emotional release also includes the freedom to slightly transgress social norms without fear of disapprobation,[93] from perusing an ex-partner's Facebook to googling embarrassing questions.[94] Big data's collection and analysis reduces the room to temporarily and minimally misbehave.

These harms are exacerbated by the lack of transparency over what data companies and governments are collecting and analyzing. In fact, some companies deliberately hide their use of big data because consumers have expressed discomfort over the intimate information companies can learn.[95] When it is unclear what data is

88. *See, e.g.*, Sanjay Srivastava et al., *Development of Personality in Early and Middle Adulthood: Set Like Plaster or Persistent Change?*, 84 J. PERSONALITY & SOC. PSYCHOL. 1041 (2003), http://www.apa.org/pubs/journals/releases/psp-8451041.pdf [https://perma.cc/WP75-F9M4].

89. *See* Westin, *supra* note 6, at 37.

90. *Cf. id.*

91. *See id.* at 38.

92. *See* Eric Deggans, *Learning How To Code-Switch: Humbling, But Necessary*, NPR (Apr. 10, 2013), http://www.npr.org/sections/codeswitch/2013/04/10/176234171/learning-how-to-code-switch-humbling-but-necessary (last visited Dec. 15, 2016).

93. Westin, *supra* note 6, at 38–39.

94. *See, e.g.*, Michael Agger, *Is It Wrong To Sleep with Your Sister?*, SLATE (Nov. 10, 2009), http://www.slate.com/articles/technology/lifehacking/2009/11/is_it_wrong_to_sleep_with_your_sister.html [https://perma.cc/EM46-QXMA].

95. *See* Duhigg, *supra* note 34.

generating which inferences, all collected information has the potential to reveal personal details and impinge autonomy.

3. Impeding Anonymity

Big data prevents people from being anonymous online. Algorithms can successfully infer people's identities even where a dataset has been scrubbed of identifying information. In one example, researchers analyzed an anonymized dataset of Netflix movie ratings and determined the identities of some of the subjects by joining movie ratings published on a different website.[96] Researchers also successfully determined the name of one of the users whose search queries were listed in an anonymized dataset of America Online searches.[97] Even short of ascertaining a user's offline identity, however, big data can construct a detailed profile of her, thereby thwarting full anonymity. A person is not fully anonymous if she is known to be a woman between the ages of 24 and 35 who lives in zip code 00501 and works in the healthcare industry.

Online anonymity forms the foundation of many of the defining characteristics of the Internet: candid discourse in comment sections, unlikely friendships, dissident action against authoritarian governments, and whistleblowing.[98] The Supreme Court has recognized a constitutional right to protecting membership lists from disclosure and authoring anonymous handbills.[99] Exposing group affiliations and author identities can chill people's willingness to join organizations and engage in expressive activities.[100] In the context of big data, inferring the identity of users online reduces people's willingness to speak, interact, and transact without inhibition.

---

96. Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 13, https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf [https://perma.cc/66A4-NNTV].

97. Michael Barbaro & Tom Zeller, *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), http://www.nytimes.com/2006/08/09/technology/09aol.html (last visited Dec. 15, 2016).

98. *See, e.g.*, Chelsea Manning, *Bradley Manning's Statement Taking Responsibility for Releasing Documents to WikiLeaks*, FREE CHELSEA MANNING (Feb. 28, 2013), https://www.chelseamanning.org/news/bradley-mannings-statement-taking-responsibility-for-releasing-documents-to-wikileaks [https://perma.cc/UZ5C-UUFC]; John Pollock, *How Egyptian and Tunisian Youth Hacked the Arab Spring*, MIT TECH. REV. (Aug. 23, 2011), https://www.technologyreview.com/s/425137/streetbook/ [https://perma.cc/QN46-85PJ].

99. *See* NAACP v. Alabama, 357 U.S. 449, 462–63 (1958); Talley v. California, 362 U.S. 60, 64–65 (1960).

100. *See Id.*

4. Eroding Belief in Human Agency

Big data's promise of inferring truths about people inflicts harm on not only data subjects but also society at large. Big data claims: with enough data and analysis, a machine can predict people's future behavior and characteristics. This proposition promulgates the fallacy that generalizations made by big data are always true, undermines belief in human agency, and aggravates latent prejudices.

The goal of big data is to generalize. Big data analyzes information for patterns, fashions those patterns into rules, and applies the rules to future data.[101] Rules that are highly specific to a set of conditions, however, are of limited value; they cannot be applied to as many future cases. This is known as overfitting.[102] To provide a discernable benefit, then, algorithms must generalize relationships between data and make inferences about people based on their match to a limited number of values. As such, big data "learns" not about an individual but about people who resemble the individual in a set of ways. Furthermore, the learning is based on using past observations to infer future instances,[103] and past performance does not guarantee future results.[104] It is true data analysis does not purport to guarantee true conclusions — under the hood, algorithms actually generate inferences of varying probabilities.[105] In application, however, its insights are treated by companies and governments as truths. Recidivism scoring software outputs quantitative risk scores of recidivism,[106] but parole boards use the scores to make binary decisions to release or detain. Colleges assign students to summer school based on a graduation score.[107] Employers hire or promote candidates according to a readiness score.[108] This increasing reliance on likelihoods to

101. *See* BRAMER, *supra* note 35, at 3.

102. *See id.* at 121.

103. *See* John Vickers, *The Problem of Induction*, *in* THE STAN. ENCYCLOPEDIA OF PHIL. (Nov. 15, 2006), http://plato.stanford.edu/entries/induction-problem (last visited Dec. 15, 2016) ("[T]he fact that the inductive habit succeeded in the past is itself only a gigantic coincidence, giving no reason for supposing it will succeed in the future.").

104. *See, e.g.*, NASSIM NICHOLAS TALEB, FOOLED BY RANDOMNESS: THE HIDDEN ROLE OF CHANCE IN LIFE AND IN THE MARKETS 108–22 (2d ed. 2005).

105. *See supra* Section III.D.

106. *See Prison*, NORTHPOINTE, INC., http://www.northpointeinc.com/solutions/prison [https://perma.cc/8CE4-BZXC].

107. Jon Marcus, *Here's the New Way Colleges Are Predicting Student Grades*, TIME (Dec. 10, 2014), http://time.com/3621228/college-data-tracking-graduation-rates/ [https://perma.cc/CZZ2-MLF4].

108. *See Cornerstone Insights*, CORNERSTONE ONDEMAND, https://www.cornerstoneondemand.com/insights [https://perma.cc/A28R-N5TQ].

make decisions propagates the misconception that big data is always able to determine truths about people.[109]

By presenting generalized and probabilistic inferences as certainties, big data creates the impression that an individual's conduct and attributes are capable of being predicted. Such an impression contradicts the proposition that humans have agency.[110] The belief that people have the capacity to make free choices and shape their personhood, however, is a core foundation of the U.S. legal system. Trust that people are not predestined to commit crimes is the reason why people are afforded the right to a fair trial and presumed innocent until proven guilty.[111] The Declaration of Independence, by proclaiming that all people have an inalienable right to life, liberty, and the pursuit of happiness, presupposes that no one is bound to a particular fate. By challenging the human capacity for free will, big data destabilizes society's adherence to democratic values.

Even if people rationally understand that an inference about a person is not necessarily fact, big data may nonetheless exacerbate the biases that exist in society. Studies show that participants subconsciously primed with racial stereotypes of blacks are more likely to rate ambiguous behavior as hostile and rate juvenile offenders as culpable.[112] In classrooms where a teacher is given less favorable information on some students, those students do not perform as well.[113] As a general matter, humans take mental shortcuts.[114] If people increasingly encounter big data generating inferences and helping to form decisions that accord with their preconceived notions of others, big data may deepen the prejudices that exist in society.

---

109. The vocabulary used to discuss big data also implies that big data can determine facts, and not mere potentials: machine learning, deep learning, predictive analytics, and information discovery.

110. It could be argued that big data does not challenge the capacity for humans to exercise free choice but rather makes observations about people who make certain choices. That proposition may be true where an algorithm uses data on a person's behavior to generate an inference about her attribute. Where big data relies on a person's attribute to make a prediction of future conduct, however, it does suggest that people's decision making is at least partly predetermined.

111. U.S. CONST. amend. VI; Coffin v. United States, 156 U.S. 432, 453 (1895).

112. *See generally* Patricia G. Devine, *Stereotypes and Prejudice: Their Automatic and Controlled Components*, 56 J. PERSONALITY & SOC. PSYCHOL. 5 (1989); Sandra Graham & Brian S. Lowery, *Priming Unconscious Racial Stereotypes About Adolescent Offenders*, 28 LAW & HUMAN BEHAVIOR 483 (2004).

113. *See generally* ROBERT ROSENTHAL, PYGMALION IN THE CLASSROOM: TEACHER EXPECTATION AND PUPILS' INTELLECTUAL DEVELOPMENT (1968).

114. *See, e.g.*, Daniel T. Gilbert & J. Gregory Hixon, *The Trouble of Thinking Activation and Application of Stereotypic Beliefs*, 60 J. PERSONALITY & SOC. PSYCHOL. 509 (1991) (finding that cognitive busyness may increase the likelihood that a person views a minority in stereotypic terms).

V. ASSESSING ALGORITHMS AND HARMS

This Note argues that the process of using data to derive inferences about people may harm rights to privacy and autonomy even if the inferences are never used to provide services or make decisions. It identifies four types of non-use harms: learning personal information, limiting autonomy, impeding anonymity, and eroding belief in human agency. However, some instances of big data do not cause these harms and yet provide valuable benefits. LinkedIn's "Jobs You May Be Interested In" feature analyzes the textual content of a profile to generate relevant job postings, and does not involve imperceptibly collecting data, inferring sensitive information about people, or using unexpected correlations.[115] Fitbit uses motion data to calculate a sleep efficiency score,[116] but the data analysis neither makes sensitive inferences nor uses unexpected correlations.[117] To differentiate between examples of big data that do and do not threaten privacy and autonomy, this Note proposes a three-step framework. The framework is based on the characteristics of big data observed in Part III and the non-use harms examined in Part IV.

(1) Does the algorithm make an inference about a person based on a correlation between data points that is unexpected?
    (a) If unexpected → go to Step (2).
    (b) If not unexpected → not harmful.
(2) Is the inferred information more sensitive than the collected information?
    (a) If more sensitive → go to Step (3).
    (b) If not more sensitive → not harmful.
(3) Does generating the inference breach contextual integrity?
    (a) If a breach of contextual integrity → harmful.
    (b) If no breach of contextual integrity → not harmful.

Step (1) asks whether an algorithm which generates an inference about a person applied a correlation that is unexpected. This Step seeks to account for the harms to privacy and autonomy that result when a person agrees to disclose some personal data and unwittingly reveals other personal information. As Section III.D discussed, this is

---

115. *See* Siya Raj Purohit, *How LinkedIn Knows What Jobs You Are Interested In*, UDACITY (May 21, 2014), http://blog.udacity.com/2014/05/how-linkedin-knows-what-jobs-you-are.html [https://perma.cc/A5FS-T3NS].

116. *How Does My Tracker Count Steps?*, FITBIT (Sept. 10, 2016), https://help.fitbit.com/articles/en_US/Help_article/1143 [https://perma.cc/X9KG-3465].

117. One of the benefits of Fitbit's sleep tracker has been helping people self-diagnose sleep apnea. *See* Kim Painter, *Sleep-Tracking Gadgets Raise Awareness — and Skepticism*, USA TODAY (Mar. 24, 2013, 9:38 AM), http://www.usatoday.com/story/news/nation/2013/03/24/sleep-tracking-devices/2007085/ [https://perma.cc/9K9E-S2NS].

more likely to occur where data analysis discovers a correlation between data that is surprising. Whether a correlation is unexpected should turn on a reasonable person standard.

This Step does not ask about the sensitivity of the information. An inference can be unexpected yet not sensitive, or expected yet highly sensitive. For example, a conclusion that a person dances ballet because she has Jet Glue and tennis balls in her gym bag involves an unexpected correlation but does not involve sensitive information.[118] Inversely, a conclusion that a person is ethnically Korean because she speaks fluent Korean involves an expected correlation but does contain sensitive information. The sensitivity of information and how it impacts whether an instance of big data is harmful is the subject of Step (2).

Step (2) asks whether the inferred information is more sensitive than the collected information that was analyzed for insights. This Step aims to capture the heightened privacy protections afforded sensitive information. The framework defines sensitivity according to the two factors discussed in Section III.C, whether the information is: (i) a qualitative conclusion as opposed to a single, quantitative data point, and (ii) related to attributes or demographic characteristics as opposed to behavior. As previously discussed, qualitative data is generally more sensitive because it pertains to who a person is rather than measurement of a single point in time; data on a person's attribute or demographic is generally more sensitive because it is tied to personhood. Generating these types of conclusions thus has a greater potential to harm privacy and autonomy interests.

Step (3) examines whether generating the inference breaches contextual norms using a version of Nissenbaum's theory modified for big data. Nissenbaum's version of contextual integrity evaluates privacy by asking whether an entity's request for or sharing of a piece of personal information comports with the norms of the context.[119] It examines the collected information in light of the context. This framework examines the inferred information in light of the context. If directly requesting the information that was inferred would have breached contextual integrity, then generating the inference breaches contextual integrity. Thus, Netflix inferring a customer's sexual orientations would breach informational norms, but a hospital generating a patient's diabetes risk score would not. Step (3) aims to preserve the insights contributed by big data by permitting analysis that generates even unexpected and sensitive conclusions so long as making the conclusions comports with contextual norms.

---

118. *See* Margaret Fuhrer, *Show and Tell: Inside Karina González's Dance Bag*, POINTE MAGAZINE (Mar. 28, 2014), http://pointemagazine.com/inside-pt/issuesaprilmay-2014show-and-tell-inside-karina-gonzalezs-dance-bag/ [https://perma.cc/WF3J-V89J].

119. *See supra* Section II.A.

VI. CONCLUSION

Big data harms privacy and autonomy interests when it is used to stultify people's beliefs, to circumvent due process, and to discriminate against groups of people. Regulating big data to mitigate these harms at the point that companies and government act on inferred information is perhaps intuitive and easy. But exclusive focus on those harms ignores the subtler ways that inferring itself threatens privacy and autonomy. The heat of an external gaze constricts the space to act without scrutiny, impedes anonymity, and undermines the belief that all humans have the capacity to shape their lives. To secure the value of big data without weakening the foundation of a flourishing society, then, regulation must target how big data both acts on us and gets to know us at all.