Printed in U.S.A.
Vol. 117, No. 2

Articles

INFORMATION PRIVACY AND THE INFERENCE ECONOMY

Alicia Solow-Niederman

ABSTRACT—Information privacy is in trouble. Contemporary information privacy protections emphasize individuals' control over their own personal information. But machine learning, the leading form of artificial intelligence, facilitates an *inference economy* that pushes this protective approach past its breaking point. Machine learning provides pathways to use data and make probabilistic predictions—inferences—that are inadequately addressed by the current regime. For one, seemingly innocuous or irrelevant data can generate machine learning insights, making it impossible for an individual to anticipate what kinds of data warrant protection. Moreover, it is possible to aggregate myriad individuals' data within machine learning models, identify patterns, and then apply the patterns to make inferences about other people who may or may not be part of the original dataset. The inferential pathways created by such models shift away from "your" data and towards a new category of "information that might be about you." And because our law assumes that privacy is about personal, identifiable information, we miss the privacy interests implicated when aggregated data that is neither personal nor identifiable can be used to make inferences about you, me, and others.

This Article contends that accounting for the power and peril of inferences requires reframing information privacy governance as a network of organizational relationships to manage—not merely a set of dataflows to constrain. The status quo magnifies the power of organizations that collect and process data, while disempowering the people who provide data and who are affected by data-driven decisions. It ignores the triangular relationship among collectors, processors, and people and, in particular, disregards the codependencies between organizations that collect data and organizations that process data to draw inferences. It is past time to rework the structure of our regulatory protections. This Article provides a framework to move forward. Accounting for organizational relationships reveals new sites for regulatory intervention and offers a more auspicious strategy to contend with the impact of data on human lives in our inference economy.

AUTHOR—Associate Professor, University of Iowa College of Law; Affiliated Fellow, Yale Law School Information Society Project; Faculty Associate, Berkman Klein Center for Internet & Society at Harvard University; Non-Resident Affiliate, Northeastern University School of Law Center for Law, Innovation and Creativity. I thank Jack Balkin, Yochai Benkler, Hannah Bloch-Wehba, Ignacio Cofone, Rebecca Crootof, Nikolas Guggenberger, Martha Minow, Richard Re, Neil Richards, Andrew Selbst, Daniel Solove, Susannah Barton Tobin, Ari Ezra Waldman, and Jonathan Zittrain for incisive comments and formative conversations. Thank you also to Evelyn Douek, Woodrow Hartzog, Thomas Kadri, Paul Ohm, Przemek Pałka, Harry Surden, and Rory Van Loo. I benefitted from the opportunity to present this project at the Climenko Fellows "Half-Baked" Workshop, the Junior Law & Tech Working Group, and the Yale ISP Fellows Writing Workshop. I am grateful to the editors of the *Northwestern University Law Review*, especially Jordana Beh, Deepa Chari, Bradford McGann, Regan Seckel, and Megha Sorot, for their excellent suggestions and editorial assistance. Any remaining errors or omissions are my own. This Article is dedicated to Nancy Solow.

INTRODUCTION

Information privacy is in trouble. Not because it's dead.[1] Not because people claim they have "nothing to hide" and do not care about it.[2] Information privacy is in trouble because the American protective regime relies on individual control over data, and machine learning (ML) stretches its underlying assumptions past their breaking point.[3] Imagine that your neighbor uploaded photographs of your housewarming party in 2010 on a social media site and "tagged" you. Several years later, a private company scrapes photographs of thousands of people from social media sites to build

[1] *See* Polly Sprenger, *Sun on Privacy: 'Get Over It!,'* WIRED (Jan. 26, 1999, 12:00 PM), https://www.wired.com/1999/01/sun-on-privacy-get-over-it [https://perma.cc/6359-NK6Q]; Judith Rauhofer, *Privacy Is Dead, Get Over It! Information Privacy and the Dream of a Risk-Free Society*, 17 INFO. & COMMC'NS TECH. L. 185, 196 n.1 (2008) (reporting the origin of the quote).

[2] *See* Ignacio N. Cofone, *Nothing to Hide, but Something to Lose*, 70 U. TORONTO L.J. 64, 64–65 (2020) (discussing the errors in the "nothing to hide" argument against privacy); Daniel J. Solove, *"I've Got Nothing to Hide" and Other Misunderstandings of Privacy*, 44 SAN DIEGO L. REV. 745, 764–72 (2007) (critiquing the "nothing to hide" response to surveillance and data mining).

[3] Here, and throughout this Article, I focus on the U.S. regulatory regime and use the term "information privacy" to refer to the "consumer protection" understanding that dominates American law, and which focuses on how private entities may collect and use personal data. The Fourth Amendment controls government data collection and use in America. *See Fourth Amendment*, EPIC.ORG, https://epic.org/issues/privacy-laws/fourth-amendment [https://perma.cc/R9K4-AHSU]. Europe's regime, by contrast, adopts a "data protection" model that controls both public and private use of data and "proceed[s] from the principle that data protection is a fundamental human right." Anupam Chander, Margot E. Kaminski & William McGeveran, *Catalyzing Privacy Law*, 105 MINN. L. REV. 1733, 1747 (2021). Because my focus is on the American regime, and because "data protection" and "data governance" are terms of art in international law that are not yet widely accepted in American law, I use the more general term "information privacy." For further description of differences between the two regimes, see *id.* at 1747–49. Moreover, unless otherwise indicated, I use the terms "data privacy" and "information privacy" synonymously.

a facial recognition tool.[4] The private company uses one of the photos from the party and, thanks to ML, generates a "faceprint" that makes it possible to take any *other* photo associated with you, online or offline, predict that it matches the faceprint in those other photos, and associate it with your name as well as any other public details about your identity. Your ability to move anonymously about the world is erased.[5] Nor are the effects limited to you. A decade after the party, a guest who happens to appear in the background of the photo is identified and arrested by a police officer as a suspect for a crime, even though that guest has never been to the state where the crime was committed.[6]

Despite the prospect of such a far-reaching impact on individuals who use platform services as well as the friends, family, and acquaintances who interact with them, there are no open-and-shut violations of information privacy regulations on the books here. Information privacy protections today, especially in the United States, center on individual control over personal information as a way to promote individual autonomy.[7] The

[4] Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Nov. 2, 2021), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [https://perma.cc/4HXU-9MLM].

[5] Others have offered trenchant critiques of facial recognition tools. For an accessible critique of facial recognition technology, see Woodrow Hartzog & Evan Selinger, *Facial Recognition Is the Perfect Tool for Oppression*, MEDIUM (Aug. 2, 2018), https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66 [https://perma.cc/9KTU-5VB4] (calling facial recognition "the most uniquely dangerous surveillance mechanism ever invented"); and Jonathan Zittrain, *A World Without Privacy Will Revive the Masquerade*, ATLANTIC (Feb 7, 2020), https://www.theatlantic.com/technology/archive/2020/02/we-may-have-no-privacy-things-can-always-get-worse/606250 [https://perma.cc/WT79-Y2MK] (detailing how surveillance technology erodes privacy rights and asserting that law should intervene because "[f]unctional anonymity is as valuable in commerce as in speech"). This Article is distinct in its use of facial recognition as a leading example of how ML data analytics affect the relationship between individuals and entities in ways that information privacy law has not adequately recognized.

[6] *See* Dave Gershgorn, *Black Teen Barred from Skating Rink by Inaccurate Facial Recognition*, VERGE (July 15, 2021, 2:37 PM), https://www.theverge.com/2021/7/15/22578801/black-teen-skating-rink-inaccurate-facial-recognition [https://perma.cc/P5GR-C2R6] (discussing the incorrect identification of a teenage girl who had never before visited the location); *see also* Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (June 24, 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html [https://perma.cc/BT8U-DX6C] (discussing the faulty facial recognition match and arrest of a Black man); Kashmir Hill, *Your Face Is Not Your Own*, N.Y. TIMES (Mar. 18, 2021), https://www.nytimes.com/interactive/2021/03/18/magazine/facial-recognition-clearview-ai.html [https://perma.cc/GG9Y-G594] (discussing the identification of a man in the background of a photograph).

[7] *See, e.g.*, Julie E. Cohen, *Turning Privacy Inside Out*, 20 THEORETICAL INQUIRIES L. 1, 3 & n.3 (2019) ("Perhaps the dominant justification for privacy is that it promotes and protects individual autonomy." (citing BEATE RÖSSLER, THE VALUE OF PRIVACY (2d ed. 2018); Anita L. Allen, *Coercing Privacy*, 40 WM. & MARY L. REV. 723, 738–40 (1999))); ARI EZRA WALDMAN, PRIVACY AS TRUST: INFORMATION PRIVACY FOR AN INFORMATION AGE 29–33 (2018) (discussing dominant literature on

underlying assumption is that regulating access to one person's data affords control over what happens with respect to that person's information privacy. But this focus on individual control and personal data covers too little because the category of information privacy is bigger than what is currently protected by the letter of the law.[8]

Contemporary information privacy protections do not grapple with the way that machine learning facilitates an *inference economy* in which organizations use available data collected from individuals to generate further information about both those individuals and about other people.[9] The inference economy trades in data through two central predictive pathways. First, ML insights about an individual can be derived from aggregations of seemingly innocuous data. When a collection of data that individuals may not even have realized they were disclosing—such as publicly available photographs or IP addresses—becomes a pathway to other information, it becomes hard to predict which bits of data are significant.[10] This result disempowers individuals who seek to shield their personal data, yet can no longer know what needs protecting.[11]

Second, developers can aggregate data about you to train a ML model that is subsequently used to make predictions about other people. Machine learning works by gathering many data points and identifying correlative

privacy as "autonomy, choice, and control"); Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1613 & n.15 (1999) (identifying "the traditional liberal understanding of information privacy, which views privacy as a right to control the use of one's personal data").

[8] Data privacy is a dynamic, rapidly changing domain. In summer 2022, after this article was finalized for publication, Congress issued a discussion draft of a proposed omnibus federal privacy bill. *See* Press Release, H. Comm. on Energy & Com., House and Senate Leaders Release Bipartisan Discussion Draft of Comprehensive Data Privacy Bill (June 3, 2022), https://energycommerce.house.gov/newsroom/press-releases/house-and-senate-leaders-release-bipartisan-discussion-draft-of [https://perma.cc/5K4X-WAB5]. The analysis that follows does not discuss that draft.

[9] I reserve further treatment of the inference economy and the manner in which it scrambles the prior understanding of the relationship among data, information, and knowledge for future work. For an early account of the relationship between information and knowledge, focused on profiling in the European context, see Mireille Hildebrandt, *Defining Profiling: A New Type of Knowledge?*, *in* PROFILING THE EUROPEAN CITIZEN: CROSS-DISCIPLINARY PERSPECTIVES 17, 29–30 (Mireille Hildebrandt & Serge Gutwirth eds., 2008). In this piece, I introduce the term "inference economy" to help crystallize the dynamics at stake for information privacy regulation today. *See infra* Part IV.

[10] *See* Steven M. Bellovin, Renée M. Hutchins, Tony Jebara & Sebastian Zimmeck, *When Enough Is Enough: Location Tracking, Mosaic Theory, and Machine Learning*, 8 N.Y.U. J.L. & LIBERTY 556, 558–59 (2014) (discussing ML's ability to "make targeted personal predictions" from the "'bread crumbs' of data generated by people," such as cell phone location data); Catherine Dwyer, The Inference Problem and Pervasive Computing 3–4 (Oct. 8, 2009) (unpublished manuscript), https://ssrn.com/abstract=1508513 [https://perma.cc/U8XC-RGJE] (offering that "[a]n inference problem occurs when someone can combine clues and pieces of information to deduce confidential information" and focusing on this concern in "pervasive computing systems").

[11] *See infra* Section II.A.

patterns among the variables.[12] Identification of these patterns is the "learning" of machine learning. An organization or entity may use these correlative patterns to classify data into groups. It then becomes possible to probabilistically infer that other individual cases are like or unlike members of the group such that a particular categorization does or doesn't apply to a third party who was not in the original dataset.[13] This result disempowers individuals about whom inferences are made, yet who have no control over the data sources from which the inferential model is generated.[14]

ML thus exposes the need to recognize two categories of data: one, personal data, and two, data that can be processed to make inferences about persons. Information privacy law today targets only the former category. Historically, statutes and regulations didn't need to cover inference generation because economic and technological limitations implicitly protected against the kinds of privacy-invasive inferential predictions that ML makes possible.[15] In the past decade, however, that baseline has shifted: the growing ease of data collection with ubiquitous sensing technologies, combined with computing advances that permit processing at previously unimagined speeds and scales, has opened new pathways for ML model development. The consequences reach beyond the "surveillance capitalis[t]" pressure to extract more data and process that "free raw material" into behavioral data that is used for commercial gain.[16] The corollary information privacy issue is the fact that the cost of extraction is falling at the same time that, thanks to ML, the potential future benefit of using the data is growing, and the ability to anticipate or understand the present day or future importance of a particular piece of data is diminishing.

---

[12] David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

[13] American law has largely failed to recognize the distinct challenges of these kinds of relationships between individuals and unrelated third parties. *See* Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573, 613–16 (analyzing the "absence of horizontal data relations in data-governance law"); *see also* JULIE E. COHEN, KNIGHT FIRST AMEND. INST. AT COLUM. UNIV., HOW (NOT) TO WRITE A PRIVACY LAW 4, https://s3.amazonaws.com/kfai-documents/documents/306f33954a/3.23.2021-Cohen.pdf [https://perma.cc/M3E9-PAGK] (critiquing privacy law's reliance on "[a]tomistic, post hoc assertions of individual control rights" that "cannot meaningfully discipline networked processes that operate at scale").

[14] *See infra* Section II.B.

[15] *See infra* Section II.B. I do not mean to suggest that this status quo was normatively ideal; rather, I underscore how the technological state of the art interacted with the legal reality, as a practical matter.

[16] Amy Kapczynski, *The Law of Informational Capitalism*, 129 YALE L.J. 1460, 1460, 1464 (2020) (defining surveillance capitalism as organizational methods "that operate[] by 'unilaterally claim[ing] human experience as free raw material for translation into behavioral data,' and processing that data to 'anticipate what you will do now, soon, and later'" (quoting SHOSHANNA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 8 (2019))).

This combination of factors constitutes the "inference economy." An inference economy would not be possible without the animating forces of surveillance capitalism and informational capitalism, which create commercial incentives to amass data often entrenched by law.[17] Yet this phenomenon is distinct, too. The full force of the inference economy depends on ML. Machine learning is a tool that, in application, changes the potential future informational value of any particular bit of data. The term inference economy underscores how ML generates information from bits of data. It also highlights how this threat to information privacy protections runs in parallel to surveillance capitalist concerns with platform firms' manipulation of user autonomy and preferences, as well as informational capitalism's concern with property law's role in facilitating the exploitation of data.

Furthermore, focusing on the social and technological dynamics of ML is useful both to better understand weaknesses in information privacy law's current approach and to forecast emerging strains on its protective regime.[18] This analysis amplifies the insights of privacy law scholars who have critiqued the current regulatory approach on many grounds, from attacking the impossibility of providing meaningful consent in the face of complex, lengthy agreements;[19] to questioning the reliance on individual rights and corporate compliance;[20] to arguing that information privacy is relational and not individualistic, in the sense that it is contingent on relationships between individuals and large technology companies[21] and among individuals themselves;[22] to contending that the traditional approach fails to account for

[17] *See* ZUBOFF, *supra* note 16, at 8; JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATION CAPITALISM 6 (2019).

[18] I do not argue that ML is wholly unique or new in revealing these challenges; rather, my point is that the social and technological dynamics of ML illuminate issues with particular force, to be taken seriously here and now. Along with a coauthor, I have adopted a similar stance in prior work. *See* Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 247 (2019) (offering that the study of AI judging "sheds light on governance issues that are likely to emerge more subtly or slowly elsewhere"); *see also* Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1885–86 (2020) (taking a similar stance).

[19] *See, e.g.*, Joel R. Reidenberg, N. Cameron Russell, Alexander J. Callen, Sophia Qasir & Thomas B. Norton, *Privacy Harms and the Effectiveness of the Notice and Choice Framework*, 11 I/S: J.L. & POL. INFO. SOC'Y 485, 490–95 (2015) (summarizing capacious literature criticizing the notice and choice system).

[20] *See, e.g.*, Ari Ezra Waldman, *Privacy, Practice, and Performance*, 110 CALIF. L. REV. 1221, 1225–26 ("[A]ll of [the privacy practices] are performative, and our acculturation to them has entrenched them and defined our relationship to, and assumptions about, privacy law.").

[21] *See, e.g.*, Neil Richards & Woodrow Hartzog, *A Relational Turn for Data Protection?*, 4 EURO. DATA PROT. L. REV. 493, 494 n.9 (2020) (compiling privacy law scholarship focused on relationships).

[22] *See, e.g.*, Solon Barocas & Karen Levy, *Privacy Dependencies*, 95 WASH. L. REV. 555, 557–58 (2020) (surveying how any one person's privacy depends on decisions and disclosures made by other people).

the scale and nature of dataflows in the digital era.[23] But these critical scholarly insights have not grappled directly with the ways in which ML can draw inferences from data and the incentives created by this potential use of data. Nor have these critiques, in the main, translated to regulatory proposals on the ground.

At present, the legislative proposals that are proliferating at the local, state, and federal level offer solutions based on an understanding of the information privacy problem that is at best incomplete. One stylized mode of intervention centers on stronger statutory protection of an individual's rights with respect to their own data. Stronger rights might be part of a regulatory package; however, individual rights to opt into or out of data collection or subsequent uses won't help if there are flaws in the individual control model to begin with.[24] Nor will the chance to opt into or out of data collection address instances such as a private company that builds its own facial recognition tool using images acquired from publicly accessible data.[25] Another stylized mode of intervention bars or constrains the use of particular kinds of technology, such as facial recognition bans or biometric regulations. Moratoria and regulatory friction may be necessary to halt immediate harms; however, they are not adaptive long-term responses and are likely to create an endless game of legislative whack-a-mole to cover the latest emerging technology.[26]

The regulatory options on the table are tactics. They operate within the same paradigm as the long-standing protective regime, centered on individual control. They are limited to expanding individual control or addressing individual technologies. Missing, still, is a strategy that accounts for who can do things with data.

---

[23] *See, e.g.*, Viljoen, *supra* note 13, at 581 ("The pursuit of user attention and uninterrupted access to dataflows amplifies forms of identitarian polarization, aggression, and even violence. Such evidence suggests that social processes of datafication not only produce violations of personal dignity or autonomy, but also enact or amplify social inequality."); COHEN, *supra* note 1716, at 6 (stating that "focusing . . . on . . . divisions threatens to diminish the underlying transformative importance of the sociotechnical shift to informationalism as a mode of development").

[24] *See* COHEN, *supra* note 13 (arguing that reliance on consent in contemporary privacy law proposals is misguided).

[25] Rachel Metz, *Anyone Can Use This Powerful Facial-Recognition Tool — And That's a Problem*, CNN (May 4, 2021, 3:53 PM), https://www.cnn.com/2021/05/04/tech/pimeyes-facial-recognition/index.html [https://perma.cc/6QDN-YLMG].

[26] For instance, despite the debate surrounding facial recognition technology, there has been little public attention to government use of other biometric technologies. *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6, 31–34 (2020) (discussing U.S. Customs and Border Protection trials of iris recognition at land borders); *see also infra* notes 195–198.

Governing information privacy in the inference economy requires addressing a distinct set of questions: which actors have the ability to leverage the data available in the world, what incentives do those organizations have, and who is potentially harmed or helped by their inferences? Answering these questions requires targeting interventions to account for the relationships between individuals and the entities that collect and process data, not merely dataflows.[27] Precise answers are imperative because the products of the inference economy are not necessarily bad. ML promises, at least in some settings, to unlock information that may help individuals left unassisted by traditional methods, such as by broadening access to medical interventions,[28] or to allow greater insight into knotty social problems, such as identifying discrimination.[29] Yet it's not always possible to predict which bits of data from which sources might be used for outcomes that retroactively seem good or bad. And it is nearly impossible for individuals to manage and respond to inferential predictions.

To gain traction on these multifaceted challenges, this Article emphasizes the importance of inferences for information privacy and underscores the distinct position of organizations that draw inferences from data. It identifies entities that collect data and entities that process that data into further information as organizational actors that occupy unique positions

[27] Other privacy scholars urge a "relational turn" in privacy law. *See, e.g.*, Richards & Hartzog, *supra* note 21, at 2 (stating that the relational turn in privacy law "looks at how the people who expose themselves and the people that are inviting that disclosure relate to each other" and is "concerned with what powerful parties owe to vulnerable parties not just with their personal information, but with the things they see, the things they can click, the decisions that are made about them"). I share Neil Richards and Woodrow Hartzog's concern that homing in on data elides critical questions of power. *Id.* at 4. This Article focuses on machine learning as a way to recenter the conversation. I contend that ML's inference economy increases the salience of organizational dynamics that have not, to date, received sustained scholarly attention. *See infra* Section II.B, Part IV.

[28] *See, e.g.*, Andrew Myers, *AI Expands the Reach of Clinical Trials, Broadening Access to More Women, Minority, and Older Patients*, STAN. UNIV. HUM.-CENTERED A.I. (Apr. 16, 2021), https://hai.stanford.edu/news/ai-expands-reach-clinical-trials-broadening-access-more-women-minority-and-older-patients [https://perma.cc/9T5K-2V9C] (reporting the potential use of AI to generate more inclusive clinical trial criteria); Tom Simonite, *New Algorithms Could Reduce Racial Disparities in Health Care*, WIRED (Jan. 25, 2021, 7:00 AM), https://www.wired.com/story/new-algorithms-reduce-racial-disparities-health-care [https://perma.cc/M3NR-Z4VJ] (reporting how AI performed better than doctors at identifying qualitative differences in MRI images of Black patients who reported knee pain).

[29] *See, e.g.*, Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 113 (2019) (suggesting algorithms can make it easier to identify discrimination); *see also* W. Nicholson Price II & Arti K. Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. 775, 778 (2021) (suggesting that machine learning can help humans to understand complex systems, such as biomedicine).

of informational power today.[30] This Article builds from the rich literature on relationality in privacy to make a complementary point: an approach that is attentive to the inference economy is not simply relational in the sense of a particular relationship between an individual and a firm, nor in the sense that my choices may affect your privacy; it is also relational in the sense of the relationships between organizations that collect data and organizations that process data to draw inferences, and how those organizations' decisions permit the application of ML models to make predictions about individuals.

These kinds of relational dynamics are more complex than what can be represented in the contemporary, control-focused approach. That approach is linear: it emphasizes dataflows between one person and one data collector. We gain descriptive purchase and prescriptive specificity when dataflows are instead situated as part of a triangle. Critically, a trilateral reframing distinguishes the task of data collection from the task of information processing and identifies which organization(s) are conducting each task.[31] Furthermore, it provides space to acknowledge that individuals may act both as subjects from whom data is collected and as objects to whom ML models are subsequently applied. And it reveals relational dependencies that represent new sites for potential interventions.

As one example, a facial recognition company such as Clearview AI can be understood as an "information processor" that scrapes and analyzes photographs obtained from "data collectors" such as Facebook, Venmo, and Google. Rather than relying solely on regulation that bans an activity, such as scraping photographs, or regulation that bans a technology, such as facial recognition, this reframing opens up other regulatory paths that focus on the nature of the relationship between the actors that handle data. Attention to these relational dynamics suggests that we may need to regulate the conduct of data collectors (here, Facebook, Google, and Venmo) in order to regulate

[30] Scholars have previously suggested the importance of thinking about privacy regulation functionally. For instance, Jack Balkin has coined the phrase "Great Chain of Privacy Being," arguing that we should categorize privacy regulations based on their place in the chain of "(1) collection of information, (2) collation, (3) analysis, (4) use, (5) disclosure and distribution, (6) sale, and (7) retention or destruction." Jack M. Balkin, *The Fiduciary Model of Privacy*, 134 HARV. L. REV. F. 11, 30 (2020). This Article is the first, to my knowledge, to argue that the activities of data collectors that amass data and information processors that draw inferences from the data they access warrant particular attention, *see infra* Section II.B, and to detail the institutional dynamics that arise by virtue of the relationship among players at different stages of data handling, *see infra* Part IV.

[31] I use the term "information processing" to refer to activities that transform data into new information that goes beyond the original data itself. *See infra* Section IV.B (discussing the shift from *data* collection to *information* processing). As used here, the term information processing is distinct from the term processing as it appears in European Union data protection law. I adopt this distinct term for conceptual specificity and reserve further study of EU law for future work.

the conduct of information processors that lack a direct relationship to the individuals whose data they collect and then use (here, Clearview AI).

This Article argues that we need a new strategic framework for information privacy protection. It proceeds in four parts. Parts I and II explain how contemporary American information privacy protections fail to anticipate or guard against ML inferences and examine the consequences of this state of affairs. Part I considers existing legal and regulatory protections, with an emphasis on the role that control over personally identifiable data in sensitive contexts plays in the protective regime. Part II brings in machine learning, first assessing how the inferential capabilities of ML route around the protections provided by the contemporary regime, and then evaluating how particular technological and economic developments have facilitated ML advances. Shifts in these economic and technological factors both disrupt implicit information privacy protections and provide enhanced inferential potential to firms and organizations with resources and incentives to develop advanced data-processing models.

Parts III and IV maintain that recent attempts to update law to contend with information privacy challenges advance solutions that do not engage with a complete understanding of the problem. Part III evaluates leading reform proposals, such as enhanced data protection laws and technological bans. These proposals, it maintains, do not provide a strategy to engage with the ways in which firms and organizations that generate ML inferences from available data are able to amass an arsenal of informational power. Part IV contends that a better strategy must account for the institutional dynamics of the inference economy. Information privacy protections are more productively understood with a triangular frame that reckons with the distinct position and power of individuals, data collectors, and information processors.[32]

Situating potential interventions within this triangular relationship is the most auspicious strategy to harness machine learning's inferential power on behalf of human beings.

[32] I am not the first to reconceptualize a linear relationship as a triangle or to suggest the payoff of a trilateral framing. *See, e.g.*, Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2014–15 (2012); Kristen E. Eichensehr, *Digital Switzerlands*, 167 U. PENN L. REV. 665, 703 & n.202 (2019) (stating that "[t]he pattern of company challenges becomes clear when the cyberspace ecosystem is understood as a triangle, composed of three separate power centers: governments, technology companies, and users" and referencing other works that discuss similar triadic framings). This Article is the first, to my knowledge, to situate the contemporary information privacy regulatory model in these terms.

I. THE LEGAL AND REGULATORY STATUS QUO

To set the stage for how and why ML strains the status quo, this Part surveys the law as it stands and offers a brief summary of the "privacy-as-control" frame, centered on notice and choice, that guides U.S. information privacy regulation. This regulatory approach emerges from a particular understanding of what privacy is and what it requires. Long-standing contestation about what privacy does or should mean notwithstanding,[33] the standard liberal understanding situates privacy as instrumental: it is necessary to protect individual autonomy.[34] Privacy is instrumental for autonomy, at a minimum, in the thin sense of securing a person's ability to determine what information about them is public or nonpublic.[35] A thicker account of autonomy positions privacy as a social value: privacy affords "breathing room" for self-determination, allowing an individual to form and re-form the self as a social being over time.[36] Thin or thick, this understanding of privacy as essential for self-definition and self-determination pervades privacy law.[37]

In order to preserve space for individual autonomy, contemporary information privacy law relies on control of information about the self.[38] Elements of this understanding trace back to Louis Brandeis and Samuel Warren's foundational 1890 law review piece, *The Right to Privacy*, which positioned privacy as the right "to be let alone."[39] Confronted with a new technology, the camera, that captured intimate moments in individual lives

---

[33] David E. Pozen, *Privacy-Privacy Tradeoffs*, 83 U. CHI. L. REV. 221, 225 (2016); *see* Jeffrey Bellin, *Pure Privacy*, 116 NW. U. L. REV. 463, 464–67 (2021).

[34] *See* Cohen, *supra* note 7, at 3 & n.3; *see also* Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1904–05 (2013) (discussing how "legal scholarship has conceptualized privacy as a form of protection for the liberal self" and exposing the flaws of this line of thinking).

[35] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 198 (1890) ("The common law secures to each individual the right of determining, ordinarily, to what extent his thoughts, sentiments, and emotions shall be communicated to others.").

[36] Cohen, *supra* note 7, at 12–13; Mireille Hildebrandt, *Privacy and Identity*, *in* PRIVACY AND THE CRIMINAL LAW 44 (Erik Claes, Antony Duff & Serge Gutwirth eds., 2006).

[37] *See* Peter Galison & Martha Minow, *Our Privacy, Ourselves in the Age of Technological Intrusions*, *in* HUMAN RIGHTS IN THE 'WAR ON TERROR' 258 (Richard Ashby Wilson ed., 2005); Viljoen, *supra* note 13, at 599–600. Salomé Viljoen notes that even for more "social" understandings of privacy grounded in thicker accounts of autonomy, "the normative basis of these arguments remains individual autonomy: datafication is wrongful, and harmful both for individuals and society, when it threatens the capacity for individuals to develop and act on their self-will." *Id.* at 602. I reserve the question of whether this conceptualization is adequate or normatively desirable, and instead make a narrower descriptive point about the version of privacy that has been most fully instantiated in American law for decades. *See* ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967) (developing privacy as value in terms of impact on individual autonomy).

[38] *See* Waldman, *supra* note 20 (manuscript at 26–29).

[39] Warren & Brandeis, *supra* note 35, at 195 (quoting THOMAS M. COOLEY, A TREATISE ON THE LAW OF TORTS OR THE WRONGS WHICH ARISE INDEPENDENT OF CONTRACT 29 (2d ed. 1879)).

and permitted popular dissemination of those snapshots in previously impossible ways, Brandeis and Warren argued that society required new protections.[40] A privacy tort, in their view, "would secure for each person the right to determine 'to what extent his thoughts, sentiments, and emotions shall be communicated to others.'"[41] This understanding of privacy focuses on preserving "a type of immunity or seclusion" for the individual.[42]

Nearly seventy years later, William Prosser structured emergent common law formulations by enumerating four privacy torts intended to protect "against emotional, reputational, and proprietary injuries."[43] These torts, along with leading theoretical accounts of privacy as "limited access to the self," focus on an "individual's desire for concealment and for being apart from others."[44] Maintaining this form of privacy is possible only if an individual has some ability to control the kinds of information that others can access about them, thereby limiting what others can do to disturb that individual.[45] This understanding of privacy as control has dominated the liberal understanding of information privacy for decades and is especially foundational in American information privacy law.[46]

Controlling access to information about the self means one thing in the village common; it means another in a globalized information age. Information becomes not only about what travels through neighbors' whisper networks, but also about what data is collected and compiled about an individual through anonymous, computerized networks. A concern with public and private entities' growing use of "automated data systems containing information about individuals" catalyzed a 1973 U.S. Department of Health, Education, and Welfare (HEW) report on Records, Computers,

---

[40] *Id.* at 195–96.

[41] Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1807 (2010) (quoting Warren & Brandeis, *supra* note 35, at 198).

[42] Daniel J. Solove, *Conceptualizing Privacy*, 90 CALIF. L. REV. 1087, 1102 (2002).

[43] Citron, *supra* note 41, at 1809. The four torts are public disclosure of private facts, intrusion on seclusion, depiction of another in a false light, and appropriation of another's image for commercial gain. *Id.* (citing William L. Prosser, *Privacy*, 48 CALIF. L. REV. 383, 422–23 (1960)).

[44] Solove, *supra* note 42, at 1102–05.

[45] *See id.* at 1110 ("The control-over-information can be viewed as a subset of the limited access conception."). I do not claim that privacy as "access" reduces to privacy as "control"; rather, by drawing this connection, I highlight the deep roots of the privacy-as-control model that undergirds information privacy, without contending that this model exhausts the universe of privacy interests. Notably, this traditional telling omits important racial components, too. *See* Anita Allen, Address at Yale ISP Ideas Lunch (May 13, 2021) (emphasizing racial and gender inequities in conceptions of privacy).

[46] Solove, *supra* note 42, at 1109–10; Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880 (2013).

and the Rights of Citizens.[47] The HEW report recommended protection of individuals' privacy interests through a proceduralized approach known as "Fair Information Practices" (FIPs).[48] This report recognized that individuals might share data with organizations yet still retain some privacy interests in that data.[49] The Code of Fair Information Practices it proposed thus sought "to both allow 'some *disclosure* of data' and afford affected individuals at least some agency in deciding 'the nature and extent of disclosure.'"[50] These principles, centered on individual control,[51] made a massive impact on privacy law across the world and ultimately set forth a general framework for information privacy.[52]

In the United States, the FIPs never translated into an overarching, generally applicable data governance statute.[53] Instead, they were operationalized through what William McGeveran has called a "consumer protection regime" that "generally allows any collection and processing of personal data, unless it is specifically forbidden."[54] This model emphasizes individuals' "notice" of, and "consent" to, the collection and use of their data.[55] The resulting "notice-and-choice" federal informational privacy regime has two main parts that complement the common law and state statutes: so-called "sectoral" statutes, and regulatory enforcement through the Federal Trade Commission (FTC).

First, sectoral statutes provide added protection in domains deemed especially sensitive, such as personal health information, credit reporting and

---

[47] ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., U.S. DEP'T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS, at viii (1973); *see* DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 49 (2015).

[48] Woodrow Hartzog, *The Inadequate, Invaluable Fair Information Practices*, 76 MD. L. REV. 952, 952, 956–57 (2017).

[49] Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1995 (2021).

[50] *Id.* (quoting ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., *supra* note 47, at 39–40).

[51] Hartzog, *supra* note 48, at 959–60 (characterizing FIPs as centered on "control over personal information" and describing the impact of this "control" conceptualization).

[52] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1734 (2010); *see also* Pam Dixon, *A Brief Introduction to Fair Information Practices*, WORLD PRIV. F. (Dec. 19, 2007), https://www.worldprivacyforum.org/2008/01/report-a-brief-introduction-to-fair-information-practices [https://perma.cc/XW9C-MZT2] (describing FIPs and how personal information data collection systems should be managed). For further detail on the history of the FIPs, including the transition from the 1973 HEW report to formal adoption by the Organisation for Economic Co-operation and Development in 1980, see Hartzog, *supra* note 48, at 957–59.

[53] A legislative proposal for an omnibus FIPs framework that would have applied to public and private entities was scaled back and applied only to federal government agencies. *See* Woodrow Hartzog & Neil Richards, *Privacy's Constitutional Moment and the Limits of Data Protection*, 61 B.C. L. REV. 1687, 1703 (2020).

[54] William McGeveran, *Friending the Privacy Regulators*, 58 ARIZ. L. REV. 959, 966 (2016).

[55] *See id.* at 978.

financial data, and educational data.[56] Congress adopted this approach in the wake of the FIPs; with this statutory turn, information privacy evolved past the common law's emphasis on redressing past harm, such as injury to feeling or reputation, and toward a forward-looking system to reduce the risk of harm to individuals.[57]

This form of privacy statute attempts to calibrate privacy protection according to the predicted level of risk.[58] First, lawmakers "identify[] a problem—'a risk that a person might be harmed in the future.'"[59] Then, they "try to enumerate and categorize types of information that contribute to the risk," with categorization both "on a macro level (distinguishing between health information, education information, and financial information) and on a micro level (distinguishing between names, account numbers, and other specific data fields)."[60]

Policymakers then prescribe particular, heightened protections for data that falls within a sensitive category, within the narrow bounds articulated by the relevant statute. For instance, the Health Insurance Portability and Accountability Act (HIPAA)'s Privacy Rule, which applies to the healthcare context, reflects a policy calculation that health information that is identified with a particular person poses enough of a risk of future harm to that individual to warrant statutory protection.[61] Once health information is

---

[56] In addition to the Freedom of Information Act of 1966 (FOIA), 5 U.S.C. § 552(a)(3)(A) (2018), and regulation of government actors via the Privacy Act of 1974, 5 U.S.C. § 552a (2018), there are several core sectoral elements. By way of example, personal health information is regulated by the Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 25, 29, and 42 U.S.C.), and associated privacy rules, 45 C.F.R. § 164.508(a) (2007). Credit reporting and financial data are addressed by the Fair Credit Reporting Act of 1970 (FCRA), 15 U.S.C. § 1681 (2018), and Title V of the Gramm-Leach-Bliley Act (GLBA), Pub. L. No. 106-102, 113 Stat. 1338 (codified at 15 U.S.C. §§ 6801–09 (2018)). Educational data is covered by the Family Educational Rights and Privacy Act of 1974 (FERPA), Pub. L. No. 93-380, 88 Stat. 484 (codified at 20 U.S.C. § 1232g (2018)). *See* Alicia Solow-Niederman, *Beyond the Privacy Torts: Reinvigorating a Common Law Approach for Data Breaches*, 127 YALE L.J. F. 614, 617–18 & n.13 (2018), https://www.yalelawjournal.org/forum/beyond-the-privacy-torts [https://perma.cc/T45L-K2DU]; *see also* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 587 (2014).

[57] Ohm, *supra* note 52, at 1733–34.

[58] *Id.* at 1734.

[59] *Id.* (quoting Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 487–88 (2006)).

[60] *Id.*

[61] This category of shielded information is known as "protected health information" (PHI). *What is PHI?*, HHS.GOV (Feb. 26, 2013), https://www.hhs.gov/answers/hipaa/what-is-phi/index.html [https://perma.cc/2AB7-X27C]. PHI is "information, including demographic information, which relates to: the individual's past, present, or future physical or mental health or condition, the provision of health care to the individual, or the past, present, or future payment for the provision of health care to the[m]," and that either "identifies the individual or for which there is a reasonable basis to believe can be used to

"deidentified," it is thought to no longer relate to an individual and, accordingly, is no longer within HIPAA's ambit.[62] In other words, for HIPAA to cover a given piece of health data, that data has to be personal in the sense of being directly linked to, and identified with, a particular person.

If a bit of health data is directly linked to a given person, then HIPAA regulations control how specified categories of persons or entities (such as doctors or hospitals, as compared to non-healthcare actors) can access or disseminate the information.[63] In this case, the individual engaging with a "covered entity" must be given notice of how the healthcare actor will make use of protected health information.[64] If the data is not connected with a particular person, then HIPAA's privacy protections do not apply, and the information can flow freely unless subject to a different statutory or regulatory restriction.[65] HIPAA and other sectoral statutes include numerous such categorizations and determinations about how to control dataflows. In the main, these statutes share a common attribute: they see the challenge as how to provide adequate opportunity to notify an individual about the collection and use of their personal data in order to control what can be done with that data and thereby preserve that same individual's informational privacy.[66]

This sectoral approach is linear: it relies on providing an individual with opportunities to control the flow of certain bits of identifiable data about them. Daniel Solove has described this approach as "privacy self-management."[67] Under this regime, "law provides people with a set of rights to enable them to make decisions about how to manage their data," and the

identify the individual." *Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, HHS.GOV (Nov. 6, 2015) [hereinafter *HIPAA PHI De-Identification*], https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [https://perma.cc/J676-PDB4].

[62] *See HIPAA PHI De-Identification*, *supra* note 61. *But see* Ohm, *supra* note 52, at 1736–38 (challenging the efficacy of deidentification to protect privacy of healthcare data).

[63] *See* Nicolas P. Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UMKC L. REV. 385, 387, 407–08 (2012); *Summary of the HIPAA Privacy Rule*, HHS.GOV (July 26, 2013), https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [https://perma.cc/G534-X77L] (explaining which actors are "covered entities" under the Privacy Rule).

[64] *See Summary of the HIPAA Privacy Rule*, *supra* note 63 (explaining that, subject to limited exceptions, a "covered entity" must provide "notice" that "describe[s] the ways in which the covered entity may use and disclose protected health information").

[65] *See* Terry, *supra* note 63, at 408; *see also* Nicolas P. Terry, *Big Data and Regulatory Arbitrage in Healthcare*, *in* BIG DATA, HEALTH LAW, & BIOETHICS 56, 59–60 (I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena & Urs Gasser eds., 2018) (discussing the limits of contemporary healthcare data protections).

[66] Hartzog, *supra* note 48, at 958–59.

[67] Solove, *supra* note 46, at 1880.

"rights to notice, access, and consent regarding the collection, use, and disclosure of personal data," in theory, permit individuals to manage their personal privacy.[68] Scholarly critique of this regime notwithstanding, this central approach has dominated federal privacy law since the 1970s.

This approach is neither limited to the federal government nor an artifact of laws that predate the digital era. Consider the California Consumer Privacy Act of 2018 (CCPA), a state law that is considered one of the leading information privacy statutes on the books in the United States.[69] Rather than adopt a sectoral approach, the CCPA takes a comprehensive tack and explicitly recognizes the importance of advanced data analytics. Specifically, the CCPA stipulates that its grant of consumer rights extends to "[i]nferences drawn from . . . [personal information] to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes."[70] The California Attorney General's first opinion interpreting the CCPA, moreover, underscores the significance of inferences as "one of the key mechanisms by which information becomes valuable to businesses," concluding that "inferences appear to be at the heart of the problems that the CCPA seeks to address."[71] The CCPA thus expands the category of information that is covered, recognizing that tools such as ML make data significant in distinct ways when it comes to personal privacy.

But this broader coverage does not represent a new strategy for how the information is regulated. Instead, the statute remains focused on individual rights. It attempts to empower individuals by providing opportunities for those individuals to obtain access to "personal information" that businesses have about them, including inferences used to "create a profile about a

---

[68] *Id.*

[69] Chander et al., *supra* note 3, at 1734. California's Attorney General approved regulations implementing the CCPA in March 2021. *See* Press Release, State of Cal. Dept. of Just., Attorney General Becerra Announces Approval of Additional Regulations That Empower Data Privacy Under the California Consumer Privacy Act (Mar. 15, 2021), https://oag.ca.gov/news/press-releases/attorney-general-becerra-announces-approval-additional-regulations-empower-data [https://perma.cc/B8BM-UWCU]. In addition, in November 2020, California voters passed a referendum, the California Privacy Rights Act (CPRA), that clarified certain consumer rights under the CCPA and created a state privacy protection agency. *See CCPA and CPRA*, INT'L ASS'N PRIV. PROS., https://iapp.org/resources/topics/ccpa-and-cpra [https://perma.cc/6CDW-AMHJ]. The CPRA takes full legal effect in January 2023, with enforcement set to begin on July 1, 2023. *See* Off. of Att'y Gen., State of Cal., Opinion No. 20-303 on the California Consumer Privacy Act 1, 8–9 (Mar. 10, 2022), https://oag.ca.gov/system/files/opinions/pdfs/20-303.pdf [https://perma.cc/6N5A-DMD5].

[70] California Consumer Privacy Act (CCPA) of 2018, CAL. CIV. CODE § 1798.140(o)(1)(K) (West).

[71] Off. of Att'y Gen., *supra* note 69, at 13.

consumer,"[72] or to opt out of data collection altogether.[73] This intervention, in the end, comes down to the same linear approach of notice, consent, and control by the affected person. Regulation of privacy remains a personal, control-centered affair.

Complementing the sectoral approach, the FTC has emerged as the leading regulator of information privacy at the federal level.[74] Recall that the U.S. legal tradition, in functional terms, positions information privacy in terms of consumer protection.[75] Whether data collection, storage, or use is seen as problematic depends not on substantive law, but rather on whether consumers have the opportunity to exercise control.[76] When the FTC examines the agreements that consumers have entered, the central questions are whether a consumer consented after the company provided notice and choice of its policies concerning consumer data, and whether that company then complied with the terms of the agreement.[77] Violations of these agreements may lead to FTC enforcement actions.

For over two decades, the FTC has used its authority under Section 5 of the FTC Act "to police unfair and deceptive trade practices" as a way to enforce private entities' privacy policies as well as other privacy statutes and transatlantic data-sharing agreements.[78] These enforcement actions do not rely on individuals' actions; rather, they target corporate conduct. Nonetheless, they reflect the same core calculation: the objective is to define privacy in terms of an individual's control over information about them, as expressed through the exercise of notice and consent rights. Applying this calculation, objectionable conduct consists of unfair or deceptive corporate practices in which consent was obtained deceptively or the collection or use

---

[72] *See, e.g.*, Off. of Att'y Gen., *supra* note 69, at 12–13 ("[I]n light of the plain meaning of section 1798.140, subdivision (o), inferences must be disclosed to the consumer upon request.").

[73] *See* Cohen, *supra* note 13 (discussing the CCPA's opt-in-or-out approach, which is codified at CAL. CIV. CODE § 1798.120(a) (West)).

[74] *See* Solove & Hartzog, *supra* note 56, at 587. State attorneys general play an important role at the state level. *See* Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 747, 748–51, 758–95 (2016). Such state-level action, as Citron notes, has potential to "fill gaps in privacy law." *Id.* at 750. Because this Article aims to foreground the gaps and liabilities of the American system as a whole, discussion of state-level regulatory enforcement is beyond its scope.

[75] *See supra* text accompanying notes 53–57.

[76] *See* Solove, *supra* note 46, at 1880 ("The goal of this bundle of [privacy] rights is to provide people with control over their personal data, and through this control people can decide for themselves how to weigh the costs and benefits of the collection, use, or disclosure of their information.").

[77] *See id.* at 1884 (describing the FTC's role as an enforcer of privacy notices). Of course, as described above, a particular sectoral statute may establish heightened protections that regulate acceptable data practices, delineate what is required to obtain consent, or impose other restrictions.

[78] Solove & Hartzog, *supra* note 56, at 585; *A Brief Overview of the Federal Trade Commission's Investigative and Law Enforcement Authority*, FED. TRADE COMM'N CONSUMER ADVICE (May 2021), https://www.ftc.gov/about-ftc/what-we-do/enforcement-authority [https://perma.cc/SC3B-FWJH].

of information violates the terms of the initial agreement.[79] This mode of enforcement can, over time, generate what Daniel Solove and Woodrow Hartzog have called a "new common law" of privacy that relies on enforcement actions and informal guidance to set forth the bounds of acceptable conduct.[80]

The FTC's "common law" approach allows the Commission to evolve by applying its control-focused regulatory approach to newly salient categories of consumer data. For example, if health-like data that is left uncovered by HIPAA becomes increasingly important, then the FTC can attempt to step into the gap. The Commission did just that in an early 2021 enforcement action involving Flo, an app designed to help women track menstruation and fertility cycles that touted the ability to "[l]og over 70 symptoms and activities to get the most precise AI-based period and ovulation predictions."[81] The FTC took action against Flo because it had shared user data with Facebook in ways that violated the app's own privacy policy.[82] Because Flo "broke its privacy promises," the company's misleading claims were subject to FTC action; thus, the Commission could use its enforcement authority to signal the realm of (un)acceptable conduct for a kind of sensitive information that was left uncovered by sectoral statutes.[83] Furthermore, recognizing the importance of this and similar data

---

[79] The FTC's deception analysis may look beyond the specific promises made in the company's privacy policy and consider the course of dealing between a consumer and the company. *See* Solove & Hartzog, *supra* note 56, at 628.

[80] *See id.* at 627.

[81] *See* FLO, https://flo.health [https://perma.cc/L8AY-N39J]. Because information of the sort that Flo gathers is collected by an app, and not in the context of a medical relationship, it is not considered healthcare data protected by HIPAA. *See* Miles Plant, *Does Your Health App Protect Your Sensitive Info?*, FED. TRADE COMM'N CONSUMER ADVICE (Jan. 13, 2021), https://consumer.ftc.gov/consumer-alerts/2021/01/does-your-health-app-protect-your-sensitive-info [https://perma.cc/X4B2-R54K].

[82] The *Wall Street Journal* first reported this development in 2019. Sam Schechner & Mark Secada, *You Give Apps Sensitive Personal Information. Then They Tell Facebook.*, WALL ST. J. (Feb. 22, 2019, 11:07 AM), https://www.wsj.com/articles/you-give-apps-sensitive-personal-information-then-they-tell-facebook-11550851636 [https://perma.cc/FVG4-2N7F]. The FTC's complaint documents these practices in detail. *See* Complaint, Flo Health, Inc., Docket No. C-4747, FTC File No. 1923133 (June 22, 2021), https://www.ftc.gov/system/files/documents/cases/flo_health_complaint.pdf [https://perma.cc/NMV7-C9QP]. The FTC settled this matter in January 2021 and issued its final decision and order in June 2021. *See* Press Release, Fed. Trade Comm'n, Developer of Popular Women's Fertility-Tracking App Settles FTC Allegations that It Misled Consumers About the Disclosure of Their Health Data (Jan. 13, 2021), https://www.ftc.gov/news-events/press-releases/2021/01/developer-popular-womens-fertility-tracking-app-settles-ftc [https://perma.cc/7WXT-2EGA]; Flo Health, Inc., Docket No. C-4747, FTC File No. 1923133 (June 22, 2021), https://www.ftc.gov/system/files/documents/cases/192_3133_flo_health_decision_and_order.pdf [https://perma.cc/2PP3-2JXG].

[83] *See* Lesley Fair, *Health App Broke Its Privacy Promises by Disclosing Intimate Details About Users*, FED. TRADE COMM'N (Jan. 13, 2021, 11:27 AM), https://www.ftc.gov/news-events/blogs/business-blog/2021/01/health-app-broke-its-privacy-promises-disclosing-intimate [https://perma.cc/MFW3-XXUH].

as health apps and connected devices become even more common features of contemporary life, the Commission is reviewing its existing regulations regarding breaches of unsecured "individually identifiable health information" that are not covered by HIPAA and has issued policy guidance clarifying the scope of its existing rule on this matter.[84]

These enforcement actions and policy stances, however, represent evolution to expand the reach of existing protections without fundamentally altering the underlying regulatory regime.[85] Such evolutionary adaptation builds from what came before, starting with the terms agreed to in the privacy policy and relying on the baseline assumption that an individual's control over their own data is central to privacy protection and exhaustive of privacy interests.

Even more innovative FTC approaches still reflect the same fundamental assumption. Consider a path-breaking FTC enforcement action to regulate the deployment of trained ML models more directly. In early 2021, the Commission entered a settlement with Everalbum, the developer of a photo-storage app called Ever.[86] The FTC's complaint alleged that the developer acted improperly when it pivoted Ever from cloud storage to facial recognition services and "deceived consumers about its use of facial recognition technology and its retention of the photos and videos of users

---

[84] *See* 16 C.F.R. § 318.2. As part of its review of the Health Breach Notification Rule, the FTC is "actively considering . . . the application of the Rule to mobile applications [like Flo] . . . that handle consumers' sensitive health information." Letter from April J. Tabor, Sec'y, Fed. Trade Comm'n, to Pam Dixon, Exec. Dir., World Priv. F. (June 17, 2021), https://www.ftc.gov/system/files/documents/cases/192_3133_-_flo_health_inc._-_comment_response_letters.pdf [https://perma.cc/L6L5-FV5N]. Moreover, in late 2021, the FTC issued a policy statement clarifying that the Rule applies to health apps and connected devices, including apps that rely on both health information (such as blood sugar) and non-health information (such as dates on a phone's calendar). FED. TRADE COMM'N, STATEMENT OF THE COMMISSION ON BREACHES BY HEALTH APPS AND OTHER CONNECTED DEVICES 1 (2021), https://www.ftc.gov/system/files/documents/public_statements/1596364/statement_of_the_commission_on_breaches_by_health_apps_and_other_connected_devices.pdf [https://perma.cc/ER3E-J6CB].

[85] This fact is unsurprising; a common law regime is, after all, incremental by nature. *See* Shyamkrishna Balganesh & Gideon Parchomovsky, *Structure and Value in the Common Law*, 163 U. PA. L. REV. 1241, 1267 (2015) (first citing P.S. ATIYAH, PRAGMATISM AND THEORY IN ENGLISH LAW (1987); then citing BENJAMIN N. CARDOZO, THE GROWTH OF THE LAW (1924); then citing OLIVER WENDELL HOLMES JR., THE COMMON LAW 1–2 (Little, Brown & Co. 1923) (1881); and then citing Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 469 (1897)); Solove & Hartzog, *supra* note 56, at 620.

[86] Complaint, Everalbum, Inc., Docket No C-4743, FTC File No. 1923172 (May 6, 2021), https://www.ftc.gov/system/files/documents/cases/everalbum_order.pdf [https://perma.cc/35QK-Z3Y8]; *see* Press Release, Fed. Trade Comm'n, California Company Settles FTC Allegations It Deceived Consumers About Use of Facial Recognition in Photo Storage App (Jan. 11, 2021), https://www.ftc.gov/news-events/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers [https://perma.cc/ULG5-3JTZ].

who deactivated their accounts."[87] The settlement is remarkable because it does more than merely require deletion of improperly collected data. It goes further, requiring Everalbum to delete any ML *model* that was trained using that data.[88] The Commission's "algorithmic disgorgement" remedy reflects a more sophisticated understanding of the fact that data matters in the applied context of ML models, and not merely at a fixed point of collection.[89] It represents the FTC's ability to adapt its remedies to reflect technological change.

Although this adaptation is pragmatic and forward-looking in some ways, it remains an evolutionary continuation of the FTC's historic approach. As with the Flo enforcement, the FTC took action because of alleged corporate deception. In the case of Everalbum, the company "represented that it would not apply facial recognition technology to users' content unless users affirmatively chose to activate the feature," but automatically applied it to users in most states; did not limit the facial recognition feature to the stated uses, and instead used images as data inputs to develop facial recognition technology; and did not comply with the company's statements that it would delete information associated with deactivated users.[90] These deceptive actions broke the promises made to users, thereby compromising individuals' ability to exercise control over their data. It was that corporate deception concerning users' control of their data disclosures that drove the FTC's enforcement action.[91]

---

[87] Press Release, *supra* note 86; Complaint, *supra* note 82.

[88] Press Release, *supra* note 86; *see* Natasha Lomas, *FTC Settlement with Ever Orders Data and AIs Deleted After Facial Recognition Pivot*, TECHCRUNCH (Jan. 12, 2021, 7:43 AM), https://techcrunch.com/2021/01/12/ftc-settlement-with-ever-orders-data-and-ais-deleted-after-facial-recognition-pivot [https://perma.cc/P24T-38G5].

[89] *See* Rebecca Kelly Slaughter, Acting Chairwoman, Fed. Trade Comm'n, Remarks at Future of Privacy Forum: Protecting Consumer Privacy in a Time of Crisis 2 (Feb. 10, 2021), https://www.ftc.gov/system/files/documents/public_statements/1587283/fpf_opening_remarks_210_.pdf [https://perma.cc/KL73-C62C] (emphasizing the "meaningful disgorgement" of "ill-gotten data" gains as an innovative remedy in privacy cases). For further discussion of why data deletion alone is insufficient for ML systems, the FTC's algorithmic-disgorgement actions to date, and a critique of the algorithmic-disgorgement remedy, see Tiffany C. Li, *Algorithmic Destruction*, SMU L. REV. (forthcoming 2022) (manuscript at 10–12, 21–24), https://papers.ssrn.com/a=4066845 [https://perma.cc/T246-9XQ2].

[90] Press Release, *supra* note 86.

[91] The FTC applied the same remedy in a settlement with WW International (formerly Weight Watchers) and its subsidiary, Kurbo, in early 2022. *See* Stipulated Order for Permanent Injunction, Civil Penalty Judgement, and Other Relief at 2, 7–8, United States v. Kurbo Inc., No. 22-cv-00946 (N.D. Cal. Mar. 3, 2022), https://www.ftc.gov/system/files/ftc_gov/pdf/wwkurbostipulatedorder.pdf [https://perma.cc/NQ58-CVJ9] (defining "Affected Work Product" to include "any models or algorithms developed in whole or in part using Personal Information Collected from Children through the Kurbo Program," and requiring destruction or deletion of such material); Press Release, Fed. Trade Comm'n, FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids'

Federal administrative enforcement thus rests on the same linear principle as sectoral statutes: both protect information privacy by regulating individuals' ability to control information about themselves.

II. MACHINE LEARNING AND INFORMATION PRIVACY PROTECTIONS

This Part details how machine learning routes around contemporary American information privacy protections and how formal legal protections have not accounted for the power of data-driven inferences or reckoned with which firms and organizations are able to wield them, and to what effect.

A. Information Privacy, Eroded

The application of machine learning technologies exposes cracks under the surface of the contemporary information privacy model. A close analysis of ML capabilities highlights two fault lines in privacy protections, which this Section explores in turn. The first involves the difficulty of determining which bits of data warrant protection: ML makes data outside of sensitive contexts far more significant, challenging any one person's ability to know what to protect. The second involves the difficulty of accounting for the reach of data-driven analysis: ML amplifies the manner in which data about one person may be used to make predictions or discern information about members of groups. Each erodes the assumptions that the contemporary regime makes about control and how an individual is positioned to exercise it.

1. The Context Challenge

The individual-centered control model of information privacy protection assumes that it's possible for a person, at the time that they are presented with a privacy policy, to assess the consequences that might flow from releasing personally identifiable data. Data analytics have long put

Sensitive Health Data (Mar. 4, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive [https://perma.cc/EQQ3-PL8N]. The FTC's complaint alleged that "Kurbo by WW" marketed its weight-loss app to children and operated with actual knowledge of its collection of personal information from children, while failing to "provid[e] direct notice . . . [and] obtain[] parents' verifiable parental consent" as required by the Children's Online Privacy Protection Act (COPPA) Rule. Complaint for Permanent Injunction, Civil Penalties, and Other Equitable Relief at 13–15, *Kurbo*, No. 22-cv-00946, https://www.ftc.gov/system/files/ftc_gov/pdf/filed_complaint.pdf [https://perma.cc/JD6A-USJ4]. The FTC's action in the WW–Kurbo settlement extends the same pattern: enforcement against a company that contravened notice-and-choice-style protections intended to control that firm's treatment of a particular category of data (here, data concerning an especially vulnerable population, children). Although this remedy is novel, its logic, wherein the enforcement action is rooted directly in a firm's violation of controlling law that emphasizes control of a specified category of data, is not.

pressure on that assumption.[92] The classic example in privacy scholarship is Target's prediction of pregnancy based on purchasing patterns, to the great chagrin of a teenager whose father became aware of her condition when he received a coupon book from the company.[93] The Target model relied on data scientist Andrew Pole's explicit identification of approximately twenty-five products "that, when analyzed together, allowed him to assign each shopper a 'pregnancy prediction' score."[94] When consumers signed up for an in-store shopping card and consented to sharing their purchasing behavior with the store, they probably didn't imagine this sort of predictive modelling.[95]

Today, the Target example is the tip of the data analytics iceberg. Imagine, for instance, a classification task, such as distinguishing photographs of Chihuahuas from photographs of blueberry muffins:[96]

[92] *See, e.g.*, Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Anonymity and Consent*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 44, 45–46 (Julia Lane, Victoria Stodden, Stefan Bender & Helen Nissenbaum eds., 2014) (underscoring, in the age of big-data analytics, "the ultimate inefficacy of consent as a matter of individual choice and the absurdity of believing that notice and consent can fully specify the terms of interaction between data collector and data subject"); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 98–109 (2014) (noting privacy problems that "go beyond just increasing the amount and scope of potentially private information" and emphasizing the challenge of "know[ing] in advance exactly when a learning algorithm will predict [personally identifiable information] about an individual," making it impossible to "predict where and when to assemble privacy protections around that data"); *see also* Katherine J. Strandburg, *Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD, *supra*, at 7, 8 & n.13 (noting widespread recognition that the "notice and consent paradigm is inadequate to confront the privacy issues posed by the big data explosion" and compiling scholarship); Dwyer, *supra* note 10, at 4 (discussing limitations of the individual-control model in the age of "pervasive computing").

[93] Kashmir Hill, *How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did*, FORBES (Feb. 16, 2012, 11:02 AM), https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did [https://perma.cc/CSW5-3AFB].

[94] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012) https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [https://perma.cc/DDT6-GF72].

[95] *See* Crawford & Schultz, *supra* note 92, at 94–95.

[96] Brad Folkens, *Chihuahua or Muffin?*, CLOUDSIGHT (May 19, 2017), https://blog.cloudsight.ai/chihuahua-or-muffin-1bdf02ec1680 [https://perma.cc/4PA6-ZPUH] (highlighting Karen Zack's delightful "Animal or Food?" Twitter thread); *see* Karen Zack (@teenybiscuit), TWITTER (Mar. 9, 2016, 7:40 PM), https://twitter.com/teenybiscuit/status/707727863571582978 [https://perma.cc/KRU4-BNV3].

FIGURE 1

How would a human perform this task? Without technology, a human being would likely identify features such as visible whiskers or the angle of the head, in the case of dogs, or paper wrappers and gooey objects streaked through the dough, in the case of muffins. Without ML technology, a programmer would need to extrapolate out from those human observations; specify attributes such as fur color, position, and pose that make a canine unlike a pastry; and code an "expert system" to make predictions based on those attributes.[97]

Now, however, ML permits a different path.[98] If provided with a sufficiently large number of photographs of Chihuahuas and photographs of

[97] This discussion in general, and the contrast between rule-based expert systems and correlational ML models in particular, is simplified for clarity. For further description of rule-based expert systems in the context of law, see Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 YALE L.J. 1957, 1959–60, 1965–68 (1990).

[98] Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 89–95 (2014) (explaining how ML classifiers can detect patterns to model complex phenomena, without explicit programming). There are many design choices to be made along the way. For an accessible discussion of all the choices that humans make in developing a ML model, from defining the problem to cleaning the data to selection of the statistical model and beyond, see Lehr & Ohm, *supra* note 12, at 669–701.

muffins, an ML algorithm can "learn" to identify patterns in the images that distinguish the two categories.[99] It does so through pathways that are distinct from human cognition: a human, for instance, might detect visible whiskers or gooey objects streaked through dough; a computer might notice certain patterns in the edges or coloration.[100] Ultimately, by exposing the training algorithm to enough data, prelabelled as "Chihuahua" or "muffin," it is possible to develop a working model that makes predictions about the right category—dog or pastry—when applied to a new image.[101]

Machine learning thus facilitates an entirely different channel through which to derive information. ML relies on detecting patterns in datasets, as opposed to making causal predictions or engaging in more formal reasoning. It's as if, rather than manually detecting patterns in purchases after asking consumers to consent to that data collection, a store collected social media posts; matched customers' names on in-store discount cards against their social media profiles; and parsed a large dataset of social media posts for grammatical and syntactical habits—such as, say, overuse of em dashes—to discern personality traits that made customers good or bad bets for a special credit card opportunity. This hypothetical is not the stuff of science fiction; indeed, one car-insurance company recently used social media text to "look for personality traits that are linked to safe driving."[102] All the store needs to do to make this scenario real is to combine a similar data-analytic approach with an internal dataset concerning which kinds of customers make for good and bad creditors. The information privacy status quo, however, doesn't account for data's amped-up analytic potential.

This explanation refers to "supervised ML," which has, to date, been the dominant method. The concerns presented here would apply with even more force to other methods of "unsupervised" and "reinforcement" learning, which require even less human involvement in training the model.

[99] For a diagram and summary of how advanced "convolutional neural networks" recognize images, see John Pavlus, *Same or Different? The Question Flummoxes Neural Networks*, QUANTA MAG. (June 23, 2021), https://www.quantamagazine.org/same-or-different-ai-cant-tell-20210623 [https://perma.cc/EN2N-2HUC].

[100] *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–98 (2018) (analyzing how ML predictions can be inscrutable and nonintuitive to humans); Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, 3–5 (2016) (discussing how ML algorithmic processes can be opaque to humans).

[101] *See* Surden, *supra* note 98, at 90–93 (describing a typical pattern detection process for detecting spam emails).

[102] *See* Graham Ruddick, *Admiral to Price Car Insurance Based on Facebook Posts*, GUARDIAN (Nov. 1, 2016, 8:01 PM), https://www.theguardian.com/technology/2016/nov/02/admiral-to-price-car-insurance-based-on-facebook-posts [https://perma.cc/U6HQ-D5UJ]; Zittrain, *supra* note 5 (drawing on the car-insurance example to emphasize how "[d]ata from one place can be used to inform another [context]").

The problem is that the linear protective regime turns on an individual's right to control data about the self. This approach relies on clear, well-delineated, nonleaky contexts for data disclosure. A consumer's mental model about how their data might be used—and hence their choice to consent to particular collection and processing—is pegged to a particular understanding of the contexts in which that data is salient.

But ML produces a context challenge. Machine-learning analytics make it practically impossible for an individual to determine how data might or might not be significant or sensitive in a future setting.[103] HIPAA is a prime example. The statute applies to healthcare data as specified in the text and associated regulations—but not to health information outside of the regulated space. Thus, nonmedical data, such as health information voluntarily offered in an online support group for individuals suffering from a particular medical condition,[104] is constrained only by, first, whether an individual had notice of and consented to the online platform's terms of service and privacy policy; and second, whether the company complied with those terms.

These stark regulatory lines do not track the ways in which data in one context might be used to discern further information about health. A post in an online group, outside of the space regulated by HIPAA, might inform a text-analysis model that predicts substance abuse.[105] Similarly uncovered by statutory protections is a category that Mason Marks calls "emergent medical data": "health information inferred by AI from data points with no readily observable connections to one's health."[106] For instance, ML analysis might

---

[103] This collapsing of context, which makes it more challenging to manage one's privacy, has a family resemblance to the concept of "context collapse" on social media networks, wherein the "flatten[ing]" of previously distinct contexts makes it more challenging to manage one's identity. *See* Alice E. Marwick & danah boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 NEW MEDIA & SOC'Y 114, 122 (2010).

[104] *See, e.g.*, Kelsey Ables, *Covid 'Long Haulers' Have Nowhere Else to Turn — So They're Finding Each Other Online*, WASH. POST (Oct. 1, 2020, 9:00 AM), https://www.washingtonpost.com/technology/2020/10/01/long-haulers-covid-facebook-support-group [https://perma.cc/97UG-HG99].

[105] Tao Ding, Warren K. Bickel & Shimei Pan, *Social Media-Based Substance Use Prediction*, ARXIV (May 31, 2017), https://arxiv.org/abs/1705.05633 [https://perma.cc/87Q3-LP75]; *see also* Emerging Technology from the arXiv, *How Data Mining Facebook Messages Can Reveal Substance Abusers*, MIT TECH. REV. (May 26, 2017), https://www.technologyreview.com/2017/05/26/151516/how-data-mining-facebook-messages-can-reveal-substance-abusers [https://perma.cc/97LG-W3CL] (discussing the Ding, Bickel & Pan study).

[106] Mason Marks, *Emergent Medical Data: Health Information Inferred by Artificial Intelligence*, 11 U.C. IRVINE L. REV. 995, 997, 1002–03 (2021); *see also* Eric Horvitz & Deirdre Mulligan, *Data, Privacy, and the Greater Good*, 349 SCIENCE 253, 253 (2015) (noting the potential for ML to make "'category-jumping' inferences about health conditions or propensities from nonmedical data generated far outside the medical context").

connect the use of religious language such as the word "pray" on Facebook to a likelihood of diabetes, or the use of particular Instagram filters to a likelihood of depression. Critically, ML approaches can generate information from data points that a disclosing party might not have even considered significant.[107] The power and peril of ML comes from the ability to discern patterns by analyzing large datasets that may be contextually unrelated.[108] Because an individual cannot predict that a particular bit of data could yield insights about sensitive matters, ML undermines the viability of relying on individual control over a protected category, such as "medical data," to shield information privacy interests. Under these conditions, it's just not feasible for the individual to predict in which spaces, and at which points, data might be relevant for processing.

This class of challenge is not limited to health information, nor to any particular sensitive setting. Return for a moment to the neighborhood big-box store. Perhaps that store uses a facial recognition tool that identifies consumers the minute they enter the store, cross-references this information to locate the person's social media profile, derives correlations about personality based on the messages posted in that profile, and then uses this profile to instruct the security officer how closely to monitor that particular shopper.[109] It seems unlikely that a social media user who consented to a platform's terms of service imagined that disclosure in that context would permit such emergent profiling. When any bit of data might be relevant in any range of future contexts, it becomes impossible for an individual to conceptualize the risks of releasing data.

To be sure, versions of this challenge existed before ML. As one analog example, if you walk in public, a passerby on the street might overhear you on a cell phone conversation confessing your ambivalence about an employment opportunity, and then turn out to be your interviewer for that job. Still, ML is a force multiplier of this latent context challenge. The

---

[107] *See* Bellovin et al., *supra* note 10, at 590–96 (detailing how different forms of ML can deduce information from large datasets).

[108] *See* Przemysław Pałka, *Data Management Law for the 2020s: The Lost Origins and the New Needs*, 68 BUFF. L. REV. 559, 592 (2020).

[109] *See* Tom Chivers, *Facial Recognition . . . Coming to a Supermarket Near You*, GUARDIAN (Aug. 4, 2019, 4:00 AM), https://www.theguardian.com/technology/2019/aug/04/facial-recognition-supermarket-facewatch-ai-artificial-intelligence-civil-liberties [https://perma.cc/M9WB-FV7K] (suggesting retail facial recognition could help to prevent shoplifting). The prospect of retail facial recognition in the United States is not, in fact, hypothetical. In 2020, Reuters confirmed that Rite Aid, over about eight years, had added facial recognition systems to two hundred stores in the United States. In New York and Los Angeles, Rite Aid had added these systems "in largely lower-income, non-white neighborhoods." Jeffrey Dastin, *Rite Aid Deployed Facial Recognition Systems in Hundreds of U.S. Stores*, REUTERS (July 28, 2020, 11:00 AM), https://www.reuters.com/investigates/special-report/usa-riteaid-software [https://perma.cc/VJV6-DH57].

technology accelerates what Margot Kaminski, drawing on work by Jack Balkin and Reva Siegel, calls "disruption of the 'imagined regulatory scene,'" which occurs when "sociotechnical change" alters "the imagined paradigmatic scenario" for a given law "by constraining, enabling, or mediating behavior, both by actors we want the law to constrain and actors we want the law to protect."[110] The deployment of ML across a range of social contexts disrupts information privacy's imagined regulatory scene. Protective regimes for information privacy disregard this reality at their peril.

2. The Classification Challenge

So, too, does ML amplify a second latent issue: the ways that data about one person may affect members of groups. Many ML models are classificatory, in the sense that they use large datasets of information about many individuals to make predictions about third parties.[111] Consider, for instance, a bevy of emerging tools that claim to predict health outcomes, including the risk that veterans will commit suicide,[112] the likelihood of onset of Alzheimer's disease,[113] the identification of conditions ranging from rare genetic disease,[114] and depression.[115] These models share a basic pattern: they require many data points, aggregate this data to form a correlation-driven model about a group, and then probabilistically infer that new cases are like or unlike members of the group such that a particular group label does or doesn't apply to those third parties.[116] The goal is typically "to make predictions or estimates of some outcome," without specifying the means to

[110] Margot E. Kaminski, *Technological 'Disruption' of the Law's Imagined Scene: Some Lessons from Lex Informatica*, 35 BERKELEY TECH. L.J. (forthcoming 2022) (manuscript at 14, 17) (citing Jack M. Balkin & Reva B. Siegel, *Principles, Practices, and Social Movements*, 154 U. PENN. L. REV. 927 (2006)).

[111] *See* Pałka, *supra* note 108, at 595 (discussing third-party externalities that flow from one person's decisions about collection of their data).

[112] Benedict Carey, *Can an Algorithm Prevent Suicide?*, N.Y. TIMES (Nov. 23, 2020), https://www.nytimes.com/2020/11/23/health/artificial-intelligence-veterans-suicide.html [https://perma.cc/Q9MN-QCJL].

[113] Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi & Melissa Naylor, *Linguistic Markers Predict Onset of Alzheimer's Disease*, ECLINICALMEDICINE, Nov. 2020, at 1, 7, https://www.thelancet.com/action/showPdf?pii=S2589-5370%2820%2930327-8 [https://perma.cc/XG8F-RPMJ].

[114] Yaron Gurovich et al., *Identifying Facial Phenotypes of Genetic Disorders Using Deep Learning*, 25 NATURE MED. 60, 63 (2019).

[115] Kyle Wiggers, *Alphabet's Project Amber Uses AI to Try to Diagnose Depression from Brain Waves*, VENTUREBEAT (Nov. 2, 2020, 12:40 PM), https://venturebeat.com/2020/11/02/alphabets-project-amber-leverages-ai-to-identify-brain-wave-data-relevant-to-anxiety-and-depression [https://perma.cc/C7E9-GTLM]; Lingyun Wen, Xin Li, Guodong Guo & Yu Zhu, *Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding*, 10 IEEE TRANSACTIONS ON INFO. FORENSICS & SEC. 1432, 1432 (2015), https://ieeexplore.ieee.org/document/7063266 [https://perma.cc/9XFF-6ADF].

[116] Each of these steps entails many human decisions. *See* Lehr & Ohm, *supra* note 12, at 669–701.

arrive at that outcome.[117] In this way, data about one person becomes part of a tool used to, in effect, make educated guesses about other people—including guesses about information that those other people might prefer not to disclose.

American information privacy law has largely failed to recognize the distinct challenges that arise when it becomes possible to make these kinds of connections between individuals (whose data is collected) and members of groups (to whom data-driven predictions are applied).[118] There is, to be sure, an increasingly active literature that emphasizes how information privacy is relational as a general matter,[119] and how big-data analytics in particular make informational privacy relational, not individual.[120] For example, Solon Barocas and Helen Nissenbaum identify the risk of a "tyranny of the minority" in big-data analytics when "the volunteered information of the few can unlock the same information about the many."[121] And more recently, Salomé Viljoen emphasizes the importance of a "relational" theory of data governance.[122] As Viljoen explains, dataflows entail not only "vertical relation[s]" between a particular individual and a data collector, but also "horizontal relations" between the individual and "others [who] share relevant population features with the data subject."[123] Viljoen focuses on the manner in which "informational infrastructures" rely on group classification to "make sense of" individuals by taking a "relevant shared feature," generating a prediction and associated "social meaning" based upon that shared feature, and then applying this prediction to a third party deemed to fall within the relevant grouping.[124]

---

[117] *Id.* at 671.

[118] *See* Viljoen, *supra* note 13, at 613 (analyzing "the absence of horizontal data relations in data-governance law"); *see also* Cohen, *supra* note 13 (critiquing privacy law's reliance on "[a]tomistic, post hoc assertions of individual control rights" that "cannot meaningfully discipline networked processes that operate at scale"). My definition of data subject covers both categories. *See infra* text accompanying notes 224–226.

[119] *See* Richards & Hartzog, *supra* note 21, at 494 n.9 (compiling important work focused on relationships in privacy law); Ian Kerr, *Schrödinger's Robot: Privacy in Uncertain States*, 20 THEORETICAL INQUIRIES L. 123, 127–32 (2019) (contending that a "relational core" is the "common denominator" of many seemingly disparate privacy theories).

[120] *See* Barocas & Nissenbaum, *supra* note 92, at 61–64.

[121] *Id.* at 61–62 (identifying this risk and citing the Target pregnancy-prediction example); *see supra* text accompanying notes 84–87.

[122] Viljoen, *supra* note 13, at 603–16; *see also* Omri Ben-Shahar, *Data Pollution*, 11 J. LEGAL ANALYSIS 104, 106 (2019) (contending that the "harms from data misuse are often far greater than the sum of private injuries to the individuals whose information is taken").

[123] Viljoen, *supra* note 13, at 607, 611–12.

[124] *Id.* at 607.

This way of understanding data relations has some resemblance to a rich transatlantic literature.[125] For instance, Brent Mittelstadt, building on the foundational work of Luciano Floridi, has advanced a theory of group privacy. Mittelstadt maintains that privacy, understood as "the right to control data about oneself," is not possible for the individual under the conditions of algorithmic classification.[126] Contending that a group or an individual's "right to inviolate personality" can be "violated when [it] is crafted externally," including through correlative, algorithmic decisional processes, he suggests that "algorithmically grouped individuals have a collective interest in how information describing the group is generated and used."[127] As Mittelstadt himself recognizes, group rights are legally and morally contested; thus, his work advances a philosophical proposal and not a policy intervention.[128]

In addition, another European scholar and frequent coauthor with Mittelstadt, Sandra Wachter, has recently focused on the legal implications of advanced data analytics.[129] Wachter argues that the European Union's data protection regime may not amply preserve privacy or protect against discrimination in the face of "affinity profiling," a data-driven online behavioral advertising practice that "looks for a similarity between the assumed interests of a user and the interests of a group."[130] Wachter

[125] For a pre-GDPR account, see, for example, David-Olivier Jaquet-Chiffelle, *Reply: Direct and Indirect Profiling in the Light of Virtual Persons*, *in* PROFILING THE EUROPEAN CITIZEN: CROSS-DISCIPLINARY PERSPECTIVES, *supra* note 9, at 34, 34–36. Jaquet-Chiffelle replies to Mireille Hildebrandt, distinguishing between individual versus group profiling by defining and discussing the difference between "direct profiling" (in which a profile is applied to the same person who provided the data) and "indirect profiling," (in which data is collected from a large population).

[126] Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHIL. & TECH. 475, 481 (2017).

[127] *Id.* at 476, 483.

[128] *See* Brent Mittelstadt, *From Individual to Group Privacy in Biomedical Big Data*, *in* BIG DATA, HEALTH LAW, & BIOETHICS, *supra* note 65, at 175, 176 (arguing that "ad hoc groups" created through big data analytics "possess privacy interests that are sufficiently important to warrant formal protection through recognition of a moral (and perhaps, in the future, legal) right to group privacy").

[129] Wachter and Mittelstadt have collaborated on several pieces concerning European law and the challenges posed by machine learning. *See generally, e.g.*, Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735 (2021) (providing recommendations to increase fairness in machine learning under European Union nondiscrimination law); Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494 (2019) [hereinafter Wachter & Mittelstadt, *Reasonable Inferences*] (advocating for a new data protection right to "help close the accountability gap currently posed by 'high-risk inferences' . . . that are privacy-invasive or reputation-damaging, or have low verifiability in the sense of being predictive or opinion-based while being used for important decisions").

[130] Sandra Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioral Advertising*, 35 BERKELEY TECH. L.J. 367, 370 (2020).

highlights the ability to draw inferences about an individual, stating that a data processor can make predictions about "[p]otentially sensitive information such as religious or political beliefs, sexual orientation, race or ethnicity, physical or mental health status, or sex or gender identity . . . from online behavior without users ever being aware."[131] This inferential capability introduces the risk of not only privacy invasions, but also "discrimination by association" if individuals are not shielded from discriminatory inferences that are drawn based on their predicted affinity with a protected group.[132] Even under stricter European data protection standards, however, inferences tend to receive "economy class" protection, at best.[133]

This work on data relations, group privacy, and associational discrimination speaks to a latent flaw in the contemporary protective regime: its focus on a particular individual's control over data about them. Data's significance, however, is not just about how that data relates to any one person. It is also about population-level group-based inferences that can be derived from individual data points. These inferences may be used to construct particular social understandings, to route around the constraints of positive privacy law, or to make classifications that are de facto linked to sensitive attributes (or to emergent categories that may not receive formal legal protections). And yet privacy law has not focused on these inferences, despite critical scholars' long-standing concern with the ability to discriminate and stereotype through data-driven analysis.[134] ML makes this oversight especially glaring: it acts as a force multiplier of these concerns. Whether seen as a difference in degree or a difference in kind, this strain on the regulatory system warrants fresh consideration.

---

[131] *Id.* at 376–77.

[132] *See id.* at 394–98. I reserve further study of affinity profiling and American antidiscrimination law for future work. For an early study of big data and discrimination in the employment context, see generally Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016), which examines the various concerns surrounding data and algorithmic techniques in the context of Title VII's prohibition of discrimination in the employment sphere.

[133] Wachter & Mittelstadt, *Reasonable Inferences*, *supra* note 129, at 56.

[134] *See* OSCAR H. GANDY JR., THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION 1 (1993); *see also* Oscar H. Gandy Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFO. TECH. 29, 30–31 (2010) ("Some of the most troublesome candidates for regulatory exclusion or control are variables that have a strong historically generated structural linkage with other measures that we have already agreed to ban[,] . . . primarily . . . the measures of socioeconomic status and attainment that are closely associated with indicators of race, ethnicity, and gender.").

* * *

Machine learning amplifies context and classification challenges that make it nearly impossible for individuals to control their data in order to control their privacy. Yet the contemporary suite of linear information privacy protections depends on individual control. The next Section suggests that economic and technological factors historically provided additional protection of information privacy interests. Shifts in these underlying, implicit constraints matter because firms and organizations that can obtain the resources to construct ML models gain the potential for enhanced inferential power that de jure information privacy protections do not address.

B. Data's Potential, Amplified

ML's challenge to the information privacy protective regime occurs because of a convergence of factors that unlock access to otherwise unavailable (or unaffordable) data, coupled with incentives to process that data and generate predictions. This section analyzes the shifts in technology and society that have allowed the power of ML-driven inferences to emerge.

1. What Machine Learning Derives from Data

ML routes around existing information privacy protections by changing the kinds of information that organizations can derive from collected data in two important senses. First, certain activities were historically too costly, too difficult, or both too costly and too difficult to accomplish. Take, by way of example, facial recognition to identify an unknown person. This task requires obtaining and aggregating configurations of biometric markers, such as the distance between a person's nose and chin and myriad other facial measurements, to make educated guesses about similar "faceprint" configurations and thereby generate an identity match.[135] Criminologists started aggregating these measurements by hand in the late nineteenth century, and over forty years ago mathematician Woody Bledsoe tried to teach a computer to use this method to match mugshots to suspects' faces.[136] But this process was hard to do in a cost-effective way.

Automation changes the calculus. And once it becomes amply efficient and affordable to use an ML technique for a task like face recognition, there

---

[135] *See* Shaun Raviv, *The Secret History of Facial Recognition*, WIRED (Jan. 21, 2020, 6:00 AM), https://www.wired.com/story/secret-history-facial-recognition [https://perma.cc/P5AW-HSR4].

[136] *Id.*; Inioluwa Deborah Raji & Genevieve Fried, *About Face: A Survey of Facial Recognition Evaluation*, ARXIV (Feb. 1, 2021), https://arxiv.org/abs/2102.00813 [https://perma.cc/A46C-4KTE]; *see also* Karen Hao, *This Is How We Lost Control of Our Faces*, MIT TECH. REV. (Feb. 5, 2021), https://www.technologyreview.com/2021/02/05/1017388/ai-deep-learning-facial-recognition-data-history [https://perma.cc/T2QZ-2QVD] (providing a more recent history of facial recognition technology).

is an eroded barrier to further, potentially privacy-invasive inferences. For instance, having located a face match, the match may then be used both to identify a person and to infer other information about the identified individual. Imagine an abusive ex-lover who posted a nude photo of their former significant other online. If that individual is walking down the street and is identified with facial recognition technology, then it is possible to, from their presence in public, connect them back to the nude photograph and, potentially, make all sorts of other inferences—warranted or not—about them. ML may accordingly enable the derivation of other kinds of information by enabling a cost-effective categorization that can then be associated with other information in the world.

In a second set of circumstances, ML's pattern-matching capabilities may themselves generate information that it was not previously possible to discern. Take, for instance, technology that purports to identify rare genetic disorders using a photograph of an individual's face.[137] Here, the technology is used to infer that the mapping of that person's facial biometrics is sufficiently similar to the faceprint of individuals with particular genetic syndromes.[138] ML may accordingly serve as more than the enabling technology: it can operate as a new kind of inferential pathway that reveals previously hidden information that is latent in an aggregated set of data.[139]

ML thus provides distinct enabling and epistemic pathways, allowing organizations and firms to infer information that people do not reveal, based on other data points.[140] The current statutory and regulatory tack does not account for the potential to draw inferences in this way.[141] By unlocking new ways that data matter in the world, ML changes what is possible for a given actor to do in a particular setting.[142] Working out what the legal

---

[137] Gurovich et al., *supra* note 114, at 60, 63.

[138] *See id.*

[139] *See* Mariano-Florentino Cuéllar & Aziz Z. Huq, *Privacy's Political Economy and the State of Machine Learning: An Essay in Honor of Stephen J. Schulhofer*, 76 N.Y.U. ANN. SURV. AM. L. 317, 328 (2021) ("[M]achine learning can be used to expand the range of data that is epistemically fruitful.").

[140] *See supra* Section II.A.

[141] *See* Michael Kassner, *Unintended Inferences: The Biggest Threat to Data Privacy and Cybersecurity*, TECHREPUBLIC (Mar. 10, 2019, 8:32 PM), https://www.techrepublic.com/article/unintended-inferences-the-biggest-threat-to-data-privacy-and-cybersecurity [https://perma.cc/S7QP-VTDX] (describing the threat of "unintended inferences" and noting the need for legal protections against them); Wachter & Mittelstadt, *Reasonable Inferences*, *supra* note 129, at 542–71 (assessing the lack of robust protection for inferences under EU law). For a discussion of the CCPA's limited exception to this general rule, see *supra* text accompanying notes 69–71.

[142] In the Fourth Amendment context, the Supreme Court has recognized how technological change affects societal privacy expectations. *See* Carpenter v. United States, 138 S. Ct. 2206, 2214 (2018) (finding that certain "digital data—personal location information maintained by a third party—d[id] not

response should be requires confronting who can exploit the technology and to what effect.

2. Who Can Capitalize on Machine Learning

The question of who can exploit data through ML models is bound up in an antecedent one: who has access to data and the means to process it into information? Technology law scholars have long observed that, when it comes to digital governance, forces beyond the law can matter at least as much as formal legal regulations. As Joel Reidenberg and Lawrence Lessig argued in the late 1990s, the digital realm is a zone of "Lex Informatica" in which regulatory constraints and affordances emerge from design choices about digital programming as much as from formal law.[143] "Code is law."[144]

Building on this understanding, Harry Surden has contended that privacy interests are protected by "*latent* structural constraints." These constraints act as "regulators of behavior that prevent conduct through technological or physical barriers in the world," and which are "by-products of the technological or physical state of the world, rather than the result of design."[145] They operate as nonlegal mechanisms that constrain conduct in ways that "reliably prohibit unwanted behavior."[146] For example, the fact that mere mortals cannot see through a wall, at least without relying on technologically mediated X-ray capabilities, operates as a latent structural constraint that reliably prohibits people from seeing into their neighbors' homes.[147] Economic factors in particular act as an essential constraint: if the "physical and technological costs imposed by the current state of the world"

---

fit neatly under existing precedents"); Riley v. California, 573 U.S. 373, 393 (2014) (asserting that analogizing a "search of all data stored on a cellphone" to searches of physical items "is like saying a ride on horseback is materially indistinguishable from a flight to the moon"); United States v. Jones, 565 U.S. 400, 416 (2012) (Sotomayor, J., concurring) (taking particular attributes of GPS monitoring into account "when considering the existence of a reasonable societal expectation of privacy in the sum of one's public movements").

143 LAWRENCE LESSIG, CODE: VERSION 2.0, at 5–7 (2006); Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 554–55 (1998); *see also* James Grimmelmann, Note, *Regulation by Software*, 114 YALE L.J. 1719, 1732–45 (2005) (assessing software as a regulatory modality).

144 LESSIG, *supra* note 143, at 5 (first citing WILLIAM J. MITCHELL, CITY OF BITS 111 (1995); and then citing Reidenberg, *supra* note 143, at 555). Under this model, law, norms, markets, and digital architecture ("code") operate as regulatory forces that can constrain "some action, or policy, whether intended by anyone or not." Lawrence Lessig, *The New Chicago School*, 27 J. LEGAL STUD. 661, 662 n.1 (1998). For further discussion of this early understanding of regulatory forces in cyberlaw, see Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 646–48 (2020).

145 Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605, 1607–08 (2007).

146 *Id.* at 1607.

147 As Harry Surden explains, this category of regulatory constraint is "conceptually similar to an initial distribution of legal entitlements" under Wesley Hohfield's formulation of rights and entitlements. *See id.* at 1608, 1611.

fall, then the constraining protection may fall alongside it.[148] Latent constraints needed to fall for ML-driven inferences to emerge. Because ML is best understood not as a fixed technology, but rather as a utility, it's more helpful to think in terms of the resources to develop it.[149] And just as generating a utility like electricity requires resources and capital—picture a turbine that requires a moving fluid and a series of blades affixed to a rotor shaft[150]—so too does ML generation require certain resource inputs.

Two especially critical ML resources are computing power and data.[151] Access to computing power, or "compute," and access to data affect privacy regulation because, as building blocks for ML tools, their scarcity or abundance determines which institutional actors can generate inferences about people. Take, first, compute. When computer scientist Alan Turing suggested that humans attempt to build intelligent machines in the 1950s,[152] his vision was not possible in part because the computers of the era did not have the hardware capability to store commands[153] and the cost of running a computer was prohibitive.[154] Computing in general, and ML in particular, progressed only with advances in hardware.[155] Much of the theory to support advanced ML techniques was actually generated in the 1980s and 1990s. Notably, although computer scientist Geoffrey Hinton began working with the now-leading method known as deep learning nearly thirty years ago,

---

[148] *Id.* at 1608 n.12.

[149] Solow-Niederman, *supra* note 144, at 655 (explaining that AI is "akin to electricity, not a lamp"). The inferences generated by an ML model are not end products on their own; rather, they must be applied in the context of a particular application or decision-making tool.

[150] *See Electricity Explained*, U.S. ENERGY INFO. ADMIN. (Nov. 1, 2021), https://www.eia.gov/energyexplained/electricity/how-electricity-is-generated.php [https://perma.cc/EH9Y-GYTF].

[151] Solow-Niederman, *supra* note 144, at 688 & n.248; *see also* Nick Srnicek, *Data, Compute, Labour*, ADA LOVELACE INST. (June 30, 2020), https://www.adalovelaceinstitute.org/blog/data-compute-labour [https://perma.cc/BTA3-B6AE] (identifying compute, data, and labor as three categories of resource needs for AI); Karen Hao, *AI Pioneer Geoff Hinton: "Deep Learning Is Going to Be Able to Do Everything,"* MIT TECH. REV. (Nov. 3, 2020), https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything [https://perma.cc/B6BR-HSE3] (reporting that the effectiveness of the now-leading ML method known as "deep learning" had long "been limited by a lack of data and computational power").

[152] *See generally* A.M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (1950) (discussing the potential for intelligent machines and exploring how machines could learn to think).

[153] Rockwell Anyoha, *The History of Artificial Intelligence*, SITNBOSTON (Aug. 28, 2017), https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence [https://perma.cc/7UR6-RWNB].

[154] *Id.*; Robert Garner, *Early Popular Computers, 1950 –1970*, ENG'G & TECH. HIST. WIKI (Jan. 8, 2018, 4:13 PM), http://ethw.org/Early_Popular_Computers,_1950_-_1970#Early_solid-state_computers [https://perma.cc/3PED-VXL6].

[155] In particular, Moore's Law, or the rule of thumb that the number of transistors that it is possible to put on a single computing chip doubles every two years, has improved processing time and driven down the cost of building more advanced computers. *See* David Rotman, *The End of the Greatest Prediction on Earth*, 123 MIT TECH. REV. 10, 10, 12 (2020).

implementing these techniques remained impossible without adequate compute.[156] In 2012, thanks to computing advances, Hinton and his graduate students brought deep-learning methods to fruition by applying the technique to classify over one million images with a historically unparalleled error rate.[157] Fast compute was necessary to unlock "neural networks" as a viable method.[158]

Critically, computing power of the necessary magnitude is not inexpensive or widely distributed. On the contrary, it is inaccessible for many public and private actors, and risks centralizing ML development in platform firms.[159] Determining what this fact means for data analysis and for information privacy protections requires accounting for another essential resource: the data itself.

All the compute in the world would not power ML unless coupled with access to adequate data.[160] That's because ML relies on access to extremely large datasets to derive patterns.[161] The past few decades have provided just such access in spades.[162] As one example, consider a form of data that is especially important in controversial ML applications: faces. It's now far easier to collect a large number of facial images in the manner required to

---

[156] *See* Hao, *supra* note 151; Cade Metz, *Finally, Neural Networks that Actually Work*, WIRED (Apr. 21, 2015, 5:45 AM), https://www.wired.com/2015/04/jeff-dean [https://perma.cc/VYD4-YDH8].

[157] Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, *in* NEURAL INFORMATION PROCESSING SYSTEMS, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 25: 26TH ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 2012, at 1097 (2012).

[158] *See* Nicholas Thompson, *An AI Pioneer Explains the Evolution of Neural Networks*, WIRED (May 13, 2019, 7:00 AM), https://www.wired.com/story/ai-pioneer-explains-evolution-neural-networks [https://perma.cc/AL58-V23M] (interviewing Geoffrey Hinton, who stated that "in the '90s . . . data sets were quite small and computers weren't that fast").

[159] *See* Solow-Niederman, *supra* note 144, at 676 & n.203; *see also* Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power*, N.Y. TIMES (Sept. 26, 2019), https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html [https://perma.cc/KT4L-AXXJ] (reporting on computer scientists' concerns that the mounting cost of AI research requires "giant data centers" and leaves "fewer people with easy access to the [requisite] computing firepower"). In a future project, *(De)Platforming Artificial Intelligence*, I plan to explore this risk of centralization in more detail.

[160] *See* Thompson, *supra* note 158 (noting the importance of both fast compute and access to data for neural networks). In theory, technological advances that require less data could abate, but not remove, this dynamic. For a discussion of the importance of technological changes in regulatory analysis, see *infra* text accompanying notes 274–278.

[161] *See* Karen Hao, *What Is Machine Learning?*, MIT TECH. REV. (Nov. 17, 2018), https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart [https://perma.cc/YN3Y-PAHW]. As Hao notes, "there are technically ways to perform machine learning on smallish amounts of data, but you typically need huge piles of it to achieve good results." *Id.* "One-shot" or "zero-shot" learning that would train ML models with less data remains by and large elusive. *See infra* note 277 and accompanying text.

[162] Kapczynski, *supra* note 16, at 1462.

develop a facial recognition tool. That's a critical shift because, as a recent survey by data scientists Deborah Raji and Genevieve Fried illustrates, access to data has long de facto regulated facial recognition attempts. Indeed, early efforts to computerize facial recognition were thwarted in part by the challenge of obtaining enough data.[163]

That's no longer the case. The push for more and more data to support deep learning increasingly led researchers to scrape the internet,[164] amassing datasets that included images from platforms such as Google Image search, YouTube, Flickr, and Yahoo News.[165] Nonetheless, even with more data from the "wild," new technical approaches remained unable to identify individuals in real-world settings, where, for example, a face might be tilted at an angle or the lighting might be dim.[166]

What changed was the fusion of data and compute, which generated mounting incentives both to collect data and to exploit available data. By 2014, researchers at Facebook had leveraged deep learning to develop a proprietary "DeepFace" model. As Raji and Fried explain, this model was "trained on an internal dataset composed of images from Facebook profile images," and reportedly labeled "four million facial images belonging to more than 4,000 identities."[167] Facebook's access to data and computing power permitted it to achieve best-in-class accuracy at a level on a par with human performance.[168] Spurred by the allure of further advances, other face datasets kept growing in size "to accommodate the growing data requirements to train deep learning models."[169] The push to commercialize the technology mounted, too. Over time, as the field became competitive and datasets continued to expand, collection techniques also shifted: in the period running from 2014 to 2019, web sources made up almost 80% of the data included in face datasets.[170] At least for sufficiently well-resourced actors, previously controlling data and compute constraints no longer apply.

---

[163] Raji & Fried, *supra* note 136.

[164] Hao, *supra* note 136 (discussing the *About Face* survey); *see also* Richard Van Noorden, *The Ethical Questions That Haunt Facial-Recognition Research*, 587 NATURE 354, 355 (2020) (reporting the growing trend, in the past decade, of scientists collecting face data without consent).

[165] Raji & Fried, *supra* note 136.

[166] *Id.*

[167] *Id.* (citing Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato & Lior Wolf, *DeepFace: Closing the Gap to Human-Level Performance in Face Verification*, 2014 IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION 1701, 1701–05, https://ieeexplore.ieee.org/document/6909616 [https://perma.cc/DBR2-LAVW]).

[168] *See id.*

[169] *Id.*

[170] Hao, *supra* note 136 (charting facial recognition data source distribution by era); *see* Raji & Fried, *supra* note 136.

3. What Data and Compute Incentivize

The shifts in facial recognition development are an example of a more generalizable pattern concerning which kinds of actors have the ability and the incentive to take advantage of data and compute resources and generate ML instruments. Initially, a lack of compute power and a lack of data prevent a particular technological method. These stopgaps serve as a constraint that, functionally, prevents intrusion on certain privacy interests.[171] Subsequently, there are pushes to amass data. In the case of facial recognition, it was the government that initially contributed to this effort.[172] In other domains, long-standing commercial drives to amass data for marketing and targeting purposes suffice to generate sufficiently large datasets.[173] In each case, the data that is collected is available in the wild or scraped despite the ostensible protection of terms of service. In each case, the relative cost of data falls because it is so readily accessible. Firms and organizations spend less and stand to gain more from data collection.

Then, the second key resource, compute, also changes. Specifically, new processing power opens up opportunities to discern inferences from this data. There is, at some time, sufficient commercial allure that a large firm internalizes and analyzes data sources[174] or obtains data aggregated by other firms.[175] As firms in other sectors see the prospect of similar gains, there are mounting incentives to acquire data and apply the technology in more spheres of life. Firms and organizations that can acquire data and afford access to compute stand to gain more, comparatively speaking, from data processing.

Changes in access to resources, in short, both erode implicit privacy protections and affect which institutional actors can leverage ML-powered inferential predictions. These problems become even more acute as it

---

[171] *See* Surden, *supra* note 145, at 1611 (describing how physical and economic facts about the world can generate a particular Hohfeldian configuration of privacy entitlements).

[172] Raji & Fried, *supra* note 136 (describing a $6.5 million government project to generate a dataset of faces consisting of images from photoshoots).

[173] *See, e.g.*, Joseph Turow, *Shhhh, They're Listening: Inside the Coming Voice-Profiling Revolution*, FAST CO. (May 3, 2021), https://www.fastcompany.com/90630669/future-of-marketing-voice-profiling [https://perma.cc/4ULP-NSC8] (warning that ML "voice profiling," using recordings from consumer calls, is the next frontier of marketing efforts).

[174] Facebook's DeepFace model epitomizes this dynamic. *See supra* text accompanying notes 167–168; *see also* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 606 (2018) (citing Strandburg, *supra* note 92, at 10) (describing Facebook's "build-it" model, which "amass[es] training data from users in exchange for a service those users want").

[175] IBM's "Diversity in Faces" dataset epitomizes this model. *See* Vance v. Amazon.com Inc., 525 F. Supp. 3d 1301, 1306 (W.D. Wash. Mar. 15, 2021) (describing how IBM obtained data from Flickr to generate a new dataset, which it then made available to other companies).

becomes harder and harder for the individuals tasked with controlling their own information privacy to discern which bits of data might be worth protecting from would-be collectors or for other, third-party individuals to anticipate how they might be affected by correlative models produced by information processors with data collected from other people.

* * *

In the face of ML's challenges to the information privacy regime, one option is inertia: let the force of the historic trajectory continue to propel us, and trust that we'll muddle through. But muddling through requires ignoring the ways in which applications of ML sustain and accelerate an inference economy. In the inference economy, the cost of data access is comparatively lower. And the potential future informational benefit that firms or organizations might realize from using ML to leverage the data at an aggregate level is comparatively higher. Data fuels the economy[176]: as we have seen, the ML-driven products of the inference economy rely on individual data. Individuals, however, cannot effectively control for the consequences of how their data is used. It's not even clear that the label "their data" identifies a coherent category in the same way.[177] The next Part contends that even would-be reformers fail to recognize the nature of the challenges that ML presents, setting the stage for Part IV's proposal for a strategic reframing.

III. THE LIMITS OF PROPOSED REFORMS

This Part argues that most of the information privacy legislative reforms on the table do not engage with the deeper question of how organizations with the capacity and resources to create ML tools are situated relative to individuals. These proposals therefore arrive at a solution that is, at best, incomplete.[178] They generally follow one of two stylized models: one,

[176] *See* COHEN, *supra* note 17, at 48 (classifying data as quasi-capital, and identifying "data flows extracted from people" as "raw material in the political economy of informational capitalism" (emphasis omitted)); *see also* Jathan Sadowski, *When Data Is Capital: Datafication, Accumulation, and Extraction*, 6 BIG DATA & SOC'Y 1, 2 (2019) (discussing "data capital"); Viljoen, *supra* note 13, at 578 (discussing data's "de facto status as quasi capital").

[177] *See supra* Section II.A.

[178] *See* JOHN DEWEY, LOGIC: THE THEORY OF INQUIRY 108 (1938) ("The way in which the problem is conceived decides what specific suggestions are entertained and which are dismissed; what data are selected and which rejected; it is the criterion for relevancy and irrelevancy of hypotheses and conceptual structures."); *see also* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1399 (2001) ("As [historian] John Dewey aptly said, 'a problem well put is half-solved.'" (quoting DEWEY, *supra*, at 108)); *cf.* ALLIE BROSH, SOLUTIONS AND OTHER PROBLEMS (2020) (suggesting the relationship between problems and solutions is not always as clear as one might expect).

generate stronger information privacy laws that continue to rely on individual control; or two, constrain the use of, or outright ban, particularly problematic kinds of technologies.

Take option one: recognize the importance of data in contemporary information privacy and expand existing legal protections through comprehensive (as opposed to merely sectoral) legislation.[179] As an illustrative set, consider the 116th Congress, which convened from January 2019 to January 2021[180] and featured a score of comprehensive (also sometimes called "omnibus") information privacy statutes[181] alongside a bevy of bills that emphasize a particular aspect of information privacy, such as personal data related to COVID-19[182] or personal information shared through digital channels.[183] Many of the proposals shift away from the traditional consumer protection model of U.S. law and toward a data protection model.[184] That step, and other shifts towards comprehensive statutes, might be a valuable tactic insofar as the problem is inadequate in breadth for sectoral regulation.

---

[179] Because this Article concerns U.S. information privacy as it is regulated at the national level, I focus my discussion on federal law and invoke state law only insofar as it is relevant to draw out the contours of the argument. The analysis that follows does not discuss the 117th Congress's draft proposal for a federal omnibus privacy bill, which was introduced after this Article was finalized for publication. *See supra* note 8.

[180] Signe Carey, *Introduction: 116th United States Congress: A Survey of Books Written by Members*, LIBR. CONG. (Aug. 26, 2020), https://www.guides.loc.gov/116th-congress-book-list [https://perma.cc/S4PM-VM6B].

[181] *See, e.g.*, Consumer Data Privacy and Security Act of 2020, S. 3456, 116th Cong. (2020); Data Broker Accountability and Transparency Act of 2020, H.R. 6675, 116th Cong. (2020); American Data Dissemination Act of 2019, S. 142, 116th Cong. (2019); Consumer Online Privacy Rights Act, S. 2968, 116th Cong. (2019); Data Care Act of 2019, S. 2961, 116th Cong. (2019); Designing Accounting Safeguards to Help Broaden Oversight and Regulations on Data Act, S. 1951, 116th Cong. (2019); Do Not Track Act, S. 1578, 116th Cong. (2019); Mind Your Own Business Act of 2019, S. 2637, 116th Cong. (2019); Online Privacy Act of 2019, H.R. 4978, 116th Cong. (2019); Privacy Bill of Rights Act, S. 1214, 116th Cong. (2019); Setting an American Framework to Ensure Data Access, Transparency, and Accountability Act, S. 4626, 116th Cong. (2019).

[182] *See, e.g.*, COVID–19 Consumer Data Protection Act of 2020, S. 3663, 116th Cong. (2020); Exposure Notification Privacy Act, S. 3861, 116th Cong. (2020).

[183] *See, e.g.*, Privacy Score Act of 2020, H.R. 6227, 116th Cong. § 2(a)(1), (b)(1) (2020) (requiring the FTC to "develop a framework for assessing the privacy practices of interactive computer services" and "develop a system for issuing a [privacy] score for an interactive computer service"); Social Media Privacy Protection and Consumer Rights Act of 2019, S. 189, 116th Cong. § 3(a)(1)(A) (2019) (requiring "the operator of [an] online platform" to provide users with ex ante notification that "personal data of the user produced during the online behavior . . . will be collected and used by the operator and third parties").

[184] Some scholars attribute this shift to the influence of the European Union's General Data Protection Regulation (GDPR). *See, e.g.*, Hartzog & Richards, *supra* note 53, at 1694 (arguing that the GDPR has " motivat[ed] European and American companies to devote significant resources to privacy and creat[ed] structures to accommodate data subjects' rights"). Others argue that it is due to the "catalyzing" effect of the California Consumer Privacy Act of 2018 (CCPA). *See* Chander et al., *supra* note 3, at 1734–37.

But these interventions, in the main, continue to rely on the same solution of individual control.[185] That tendency may reflect economic motives,[186] political dynamics,[187] sheer inertia,[188] or some complex combination of these and other factors. Regardless of the root cause, as Julie Cohen emphasizes, the reality is that "[m]ost of the bills introduced in the 116th Congress begin by assigning sets of control rights to consumers."[189] They embrace new versions of the same approach, rooted in individual consent to data collection and processing—despite an extensive literature detailing the problems with that tack.[190] The dominant approach does not engage with the challenge of identifying who has amassed data and who has the capacity to make inferences with data, and to what effect. Bills that incorporate other tactics, such as imposing duties of loyalty on online service providers, are outliers.[191]

Proponents of option two—constrain or ban the use of particular technologies or the use of particular categories of data—come closer to grappling with which entities have the power to affect individuals in the inference economy. As one example, consider proposed or enacted bans on

---

[185] *See supra* Part I.

[186] *See, e.g.*, Hartzog, *supra* note 48, at 959 (suggesting that the FIPs likely took root for economic reasons and arguing that "[t]he 'control' conceptualization of privacy is built for globalization of the data trade").

[187] *See, e.g.*, PRISCILLA M. REGAN, LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY 177–86 (1995) (examining Congress, privacy, and policy decisions, arguing that "[p]rivacy as an idea has not had a powerful influence on policy making" and assessing the role of interests in information privacy policymaking).

[188] For discussion of the long-standing centrality of this control-centered model, see *supra* Part I.

[189] COHEN, *supra* note 13, at 4; *see also* Waldman, *supra* note 20 (manuscript at 30) (documenting how even seeming innovations in recent proposed statutes continue to "reflect long-standing privacy-as-control discourse and practices"). For a discussion of the notice-and-choice regime and how it relies on individual control, see *supra* Part I.

[190] *See, e.g.*, Solove, *supra* note 46, at 1882–93 (arguing that "privacy self-management faces several problems that together demonstrate that this paradigm alone cannot serve as the centerpiece of a viable privacy regulatory regime").

[191] Senator Brian Schatz's Data Care Act, for instance, would impose duties of care, loyalty, and confidentiality on online-service providers that collect and process "individual identifying data" from users. The bill also includes provisions to extend those duties to third-party organizations with whom an online-service provider shares the covered data. *See* Data Care Act of 2019, S. 2961, 116th Cong. § 3 (2020). Putting to the side debates concerning the viability and wisdom of such duties, *see, e.g.*, Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV. L. REV. 497, 501 (2019) (arguing that an information fiduciary model both contains internal weaknesses that it cannot resolve and raises other problems), and even assuming arguendo that they are a good solution, they are not a silver bullet. Briefly, such a set of fiduciary-inspired duties relies on an explicit agreement between the initial user and the initial data collector as well as a relationship between the entity that is collecting the data and the entity that is processing the data. But these baseline conditions do not map neatly onto all of the organizational relationships in the inference economy. For more detailed analysis of the more complex relationships at play, what such duties might (not) do with respect to regulating inferences, and how such duties might be part of a regulatory toolkit, see *infra* Section IV.B.

the use of facial recognition technology.[192] Responding in part to mounting evidence that facial recognition systems give higher false-positive rates for people of color, a growing movement aims to ban the use of the technology. A growing number of local, state, and federal legislators have proposed or enacted regulations to limit use. Especially insofar as these technologies have inequitable racial effects[193] or are used by law enforcement officers in ways that contravene best practices,[194] such bans may be a much needed policy intervention.

---

[192] For a summary of state legislation, see Nicole Sakin, *Will There Be Federal Facial Recognition Regulation in the US?*, IAPP (Feb. 11, 2021), https://iapp.org/news/a/u-s-facial-recognition-roundup [https://perma.cc/N76V-JTZS]. In addition, in June 2021, Maine enacted a statute strictly regulating facial recognition use by public employees and public officials. *See* Jake Holland, *Maine Law Curtails Facial Recognition Use by Government, Police*, BLOOMBERG L. (July 1, 2021, 11:16 AM), https://news.bloomberglaw.com/tech-and-telecom-law/maine-law-curtails-facial-recognition-use-by-government-police [https://perma.cc/347U-H2SW]. For a summary of enacted local regulations, see Nathan Sheard & Adam Schwartz, *The Movement to Ban Government Use of Face Recognition*, ELEC. FRONTIER FOUND. (May 5, 2022), https://www.eff.org/deeplinks/2022/05/movement-ban-government-use-face-recognition [https://perma.cc/AP44-MLGW]. For a discussion of enacted and proposed local regulations, see Susan Crawford, *Facial Recognition Laws Are (Literally) All Over the Map*, WIRED (Dec. 16, 2019, 8:00 AM), https://www.wired.com/story/facial-recognition-laws-are-literally-all-over-the-map [https://perma.cc/8MZS-DSEW].

For a discussion of state-level initiatives, see Pollyanna Sanderson, *Privacy Trends: Four State Bills to Watch That Diverge from California and Washington Models*, FUTURE PRIV. F. (May 25, 2021), https://fpf.org/blog/privacy-trends-four-state-bills-to-watch-that-diverge-from-california-and-washington-models [https://perma.cc/BCS9-FZRB], and Pam Greenberg, *Spotlight | Facial Recognition Gaining Measured Acceptance*, NAT'L CONF. STATE LEGISLATURES (Sept. 18, 2020), https://www.ncsl.org/research/telecommunications-and-information-technology/facial-recognition-gaining-measured-acceptance-magazine2020.aspx [https://perma.cc/335W-NYTR]. Nearly 40% of states considered bills to limit use of biometric technologies by government or by commercial entities in the year 2020, and Washington regulates facial recognition by government actors. *Id.* For a proposed federal statute to limit law enforcement use of facial recognition, introduced by Senator Edward Markey, see Facial Recognition and Biometric Technology Moratorium Act of 2020, S. 4084, 116th Cong. (2020).

[193] *See* Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. TIMES (Jan. 6, 2021), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html [https://perma.cc/9KUZ-8E79] (reporting on the third known instance of a Black man wrongfully arrested based on an inaccurate facial recognition match); *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, NAT'L INST. STANDARDS & TECH. (May 18, 2020), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software [https://perma.cc/QJG6-Z9MZ] (documenting high rates of false positives for Asians, African Americans, and Native groups in a set of 189 facial recognition algorithms); *see also* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, PROC. MACH. LEARNING RSCH.: CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, 2018, at 1, 12 (documenting higher error rates in the application of facial recognition tools to darker-skinned women).

[194] *See* Clare Garvie, *Garbage in, Garbage out: Face Recognition on Flawed Data*, GEO. L. CTR. ON PRIV. & TECH. (May 16, 2019), https://www.flawedfacedata.com [https://perma.cc/5SCF-JHTU] (reporting that one New York detective decided that a suspect resembled an actor; looked up the actor on Google to obtain high-quality images; and then used images of the actor in lieu of the suspect's face,

Regardless of one's stance on the merits, the tactic of constraining the use of a particular kind of technology is not a strategy for information privacy protection as a whole. Such a solution frames the problem in terms of how to use law to prevent a technological outcome that is deemed undesirable. This technology-versus-law showdown raises its own set of challenges. In practical terms, a "tech-specific" move to "regulate a technology rather than the conduct it enables" may quickly "become irrelevant with the advent of a newer technology not covered by the law."[195] With technologies such as iris recognition already reportedly in use by U.S. Customs and Border Protection,[196] not to mention emerging gait-recognition instruments that could also pick up on information available whenever anyone steps out in public,[197] there is a risk of whack-a-mole as legislators update law to account for rapid diffusion of technologies with similar risk profiles to facial recognition,[198] likely after they have already caused harms that direct public attention to the technology. At bottom, a ban is a political solution: it may succeed when there is adequate bottom-up mobilization or concern about specific, articulated harms to enact a particular measure, but it is not a broader legal strategy for a protective regulatory regime.

---

resulting in a "match" for a suspect whose own face had not turned up any results); Alfred Ng, *Police Say They Can Use Facial Recognition, Despite Bans*, MARKUP (Jan. 28, 2021, 8:00 AM), https://themarkup.org/news/2021/01/28/police-say-they-can-use-facial-recognition-despite-bans [https://perma.cc/7PE4-MS4K] (describing cases in which law enforcement officers failed to disclose their use of facial recognition technology in their police reports).

[195] Rebecca Crootof & BJ Ard, *Structuring Techlaw*, 34 HARV. J.L. & TECH. 347, 368 (2021); *see also id.* at 412 (noting the risk that tech-specific laws will "create legal gaps and underinclusive rules").

[196] *See* ENGSTROM ET AL., *supra* note 26, at 31–34; *see also* Press Release, Iris ID, Iris ID Products Implemented at US-Mexico Border Crossing (Jan. 19, 2016), https://www.irisid.com/iris-id-products-implemented-at-us-mexico-border-crossing [https://perma.cc/XG99-FC7X] (reporting a pilot program to test iris scanning to identify noncitizens at the U.S.–Mexico land border).

[197] *See* ROYAL SOC'Y, FORENSIC GAIT ANALYSIS: A PRIMER FOR COURTS 28 (2017) (discussing biometric gait analysis). At least some EU constituencies have expressed concern with the use of any biometric surveillance technologies in public spaces. *See* Press Release, Eur. Data Prot. Bd., EDPB & EDPS Call for Ban on Use of AI for Automated Recognition of Human Features in Publicly Accessible Spaces, and Some Other Uses of AI That Can Lead to Unfair Discrimination (June 21, 2021), https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_en [https://perma.cc/P8VD-GBEH] (calling for "a general ban on any use of AI for automated recognition of human features in publicly accessible spaces, such as recognition of faces, gait, fingerprints, DNA, voice, keystrokes and other biometric or behavioural signals, in any context").

[198] Some proposed legislation regulates biometric data more generally. *See, e.g.*, Facial Recognition and Biometric Technology Moratorium Act of 2020, S. 4084, 116th Cong. (2020) (proposing a ban on use of specified biometric systems, such as facial recognition, gait recognition, and voice recognition, by federal or state government actors). Because this proposal would not apply to commercial uses of the technology or local government actors, however, it leaves a broad swath of uses uncovered and does not contend with the relationships between data collectors and data subjects in the commercial context. For a different framing, see *infra* Part IV.

Although some legislative interventions operate one level up and regulate the category of data involved in facial recognition—biometric data—these regulations embrace the individual-control model and run into the same kinds of limitations as other interventions.[199] Illinois's Biometric Information Privacy Act, a leading state statute, is a case in point.[200] Constraining categories of a technology or categories of data advances a solution to one set of problems, pegged to that particular technology or category of data. It might be a smart tactic. But especially when it continues to rely on a linear control approach to information privacy protections, it still does not move closer to a strategy that accounts for who has the capacity or incentive to draw inferences from data. The next Part proposes a different approach.

IV. ACCOUNTING FOR THE INFERENCE ECONOMY

The inference economy imbues those who can collect data and those who can process data into information with power. These entities obtain informational power because of the inferences that they can make about individuals. ML is the leading technological engine to generate the information that gives firms and organizations power. To respond to these dynamics, this Part argues that we need to focus attention on the relationships between individuals and entities that leverage data, and not on individual control of data itself. The inference economy is not a problem to be solved; it is a reality to which to adapt. The most auspicious approach is to understand data privacy dynamics as a triangle that consists of data collectors, information processors, and individuals.

A. Recognizing Inferential Power

The information society. The information age. Surveillance capitalism. Informational capitalism. These designations recognize a monumental shift away from an economy driven by industrial capitalism and towards an economy driven by information. But beneath the headline phrase, the specific changes in resources that catalyze these shifts often get overlooked. Informational capitalism, for instance, depends on a relationship between

[199] Some European proposals avoid this problem by calling for more general bans. *See, e.g.*, Press Release, *supra* note 197. This proposal is grounded in EU legal understandings of data protection as a fundamental right, which is distinct from the American consumer protection approach to information privacy. *See supra* note 3. I reserve further analysis of the EU's AI proposed regulatory package for future work.

[200] 740 ILL. COMP. STAT. 14/15(d) (2008) (requiring individual opt-in for biometric information or identifiers); *see* Cohen, *supra* note 13, at 4 (describing the Biometric Information Privacy Act as "adopt[ing] a control-rights-plus-opt-in-or-out approach").

data, information, and knowledge, structured by law.[201] In Manuel Castells's formulation of informational capitalism, information is understood as "data that have been organized and communicated."[202] As we have seen, ML's inferential capabilities change the ways that data can be organized to produce value within a society. Indeed, in a world where anything in public is, in theory, potential fuel for an algorithm, ML changes what it means to communicate information in the first instance.

Shutting down dataflows wholesale in response to such changes is neither feasible nor socially desirable.[203] Data is not something to be stopped at the level of the individual; rather, the challenge is structuring interventions to check data's power. Individual rights have their place. But individual control, as we have seen, is a flawed privacy-protection paradigm. A complementary structural strategy is needed.

Confronting data's power at a structural level requires accepting that good and bad uses are not self-defining. The informational products of the inference economy can help as well as harm. Even outside of concerns about surveillance capitalism or behavioral manipulation through advertising, there are compelling harms. One leading instance is when ML-driven tools serve up predictive results that have a disparate impact on already-marginalized populations when applied to distribute benefits or impose burdens. Facial recognition instruments are a prime example; these

---

[201] COHEN, *supra* note 1716, at 48 ("[D]ata flows extracted from people play an increasingly important role *as raw material* in the political economy of informational capitalism."); *see also id.* at 5–6 (citing Castells's definition of informational capitalism).

[202] MANUEL CASTELLS, THE RISE OF THE NETWORK SOCIETY 17 n.25 (2d ed. 2010) (quoting MARC URI PORAT, OFF. OF TELECOMMS., THE INFORMATION ECONOMY: DEFINITION AND MEASUREMENT 2 (1977)).

[203] Some proposed interventions may face First Amendment challenges. Free speech's relationship to privacy is "long and complicated." Neil M. Richards, *Why Data Privacy Law Is (Mostly) Constitutional*, 56 WM. & MARY L. REV. 1501, 1504 (2015). Despite this messy relationship, there is enough unsettled that it is a mistake to use the First Amendment to foreclose a debate about what forms of public regulation are optimal for the inference economy. For instance, whether particular data-driven processes are speech in the first instance, and whether their regulation is able to withstand judicial scrutiny, remains an open question. *See* Brief of Amici Law Professors in Opposition to Defendant's Motion to Dismiss at 2, ACLU v. Clearview AI, Inc., No. 2020-CH-04353 (Ill. Cir. Ct. Nov. 2, 2020) (arguing that Clearview AI's facial-analysis technique is best understood as an "industrial process" that does not implicate speech rights (citing Patel v. Facebook, Inc., 932 F.3d 1264, 1268 (9th Cir. 2019))). That's especially true because different kinds of information practices, such as data collection versus analysis versus use, raise different kinds of First Amendment considerations. Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1194 (2016) (citing Neil M. Richards, *Reconciling Data Privacy and the First Amendment*, 52 UCLA L. REV. 1149, 1181–82 (2005)). Precisely because of the complexity of the relationship and the need for careful analysis of the kind of regulation at issue and the work that underlying theories of the First Amendment do in reconciling any tension, it's too hasty to assert that any regulation that affects what an actor may or may not do with data is unconstitutional. Whether or not a given intervention affects protected speech at all depends on careful, context-specific analysis, as well as the details of how a regulation is tailored.

technologies are now discussed at least as much in the language of civil rights as in privacy discourse.

This sort of harm can be subtler, too. Take, for instance, a hospital's turn to ML to determine which patients should receive extra medical care.[204] The hope was that identifying patterns in medical data would allow the hospital to make better predictions about health needs.[205] Instead, the running model's "prediction on health needs [wa]s, in fact, a prediction on health costs."[206] And because it relied on past healthcare spending to assess need, and because, "[a]t a given level of health," Black patients generated lower costs than white patients, the running model undercounted the actual needs of Black patients.[207] The inferential patterns drawn from the data in this healthcare setting hurt those most in need. In other healthcare settings and in many other instances, the applied use of ML risks embedding long-standing racial and socioeconomic inequity.[208]

The relationship between equity and ML is not quite so simple, though. Tools can also expose hidden discrimination in social systems. Racial inequity in healthcare is one such problem.[209] Racial disparities in physician assessment of pain are a well-known example.[210] In particular, knee osteoarthritis disproportionately affects people of color, yet traditional measurement techniques tend to miss physical causes of pain in these populations.[211] To counter this outcome, a research team developed a new ML approach that is able to scan X-rays and better predict actual patient pain.

---

[204] *See* Carolyn Y. Johnson, *Racial Bias in a Medical Algorithm Favors White Patients over Sicker Black Patients*, WASH. POST (Oct. 24, 2019, 2:00 PM), https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients [https://perma.cc/WD4R-A3G6].

[205] Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 336 SCIENCE 447, 449 (2019).

[206] *Id.*

[207] *Id.* at 449–50.

[208] This risk is especially acute in the criminal justice context. *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2224 (2019) (assessing racial inequality in algorithmic risk assessment); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1045 (2019) (assessing how algorithmic criminal justice affects racial equity).

[209] *See* William J. Hall, Mimi V. Chapman, Kent M. Lee, Yesenia M. Merino, Tainayah W. Thomas, B. Keith Payne, Eugenia Eng, Steven H. Day & Tamera Coyne-Beasley, *Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review*, 105 AM. J. PUB. HEALTH e60, e61 (2015).

[210] Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt & M. Norman Oliver, *Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences Between Blacks and Whites*, 113 PNAS 4296 (2016), https://www.pnas.org/doi/epdf/10.1073/pnas.1516047113 [https://perma.cc/B46M-DWZB].

[211] Emma Pierson, David M. Cutler, Jure Leskovec, Sendhil Mullainathan & Ziad Obermeyer, *An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations*, 27 NATURE MED. 136, 136 (2021).

This applied use of ML narrowed health inequities by deriving inferential patterns that help those most in need.[212]

The valence of still other cases, moreover, is mixed. Another healthcare example is illustrative: the use of ML to analyze more than three million Facebook messages and over 140,000 Facebook images and predict "signals associated with psychiatric illness."[213] This study revealed, for instance, that individuals with mood disorders tend to post images with more blues and fewer yellows, and that "netspeak" such as "lol" or "btw" was used much more by individuals with schizophrenia spectrum disorder.[214] On the one hand, such insights might identify individuals with psychiatric illness earlier, thereby helping them to obtain early intervention services associated with better outcomes. On the other hand, limiting the impact of such insights to "consented patients receiving psychiatric care" is likely to be more difficult than the researchers anticipate.[215] For example, a firm might arrive at similar results if it instead relied on statements about illness made in another setting, such as an online discussion group to support individuals, and paired the statements with social media photographs.[216] Granted, that model could not be said to have clinical validity, but it would still permit a prediction of status based on the correlation of bits of data—images and text—that were shared in a different context.[217] Moreover, the result holds the potential to affect third parties to whom the model is applied to make predictive inferences about them, even if these third parties have not made any public statements about their illness.[218] Reasonable minds can differ on whether, on net, this outcome is good or bad.

Asking whether a tool helps or harms is the wrong question. The better set of questions is: who does the tool purport to help, with what costs, and how are the costs and benefits distributed?[219] The next Section offers a restructured framework for understanding these dynamics and more effectively tailoring interventions.

---

[212] *Id.* at 136–37, 139.

[213] Michael L. Birnbaum, Raquel Norel, Anna Van Meter, Asra F. Ali, Elizabeth Arenare, Elif Eyigoz, Carla Agurto, Nicole Germano, John M. Kane & Guillermo A. Cecchi, *Identifying Signals Associated with Psychiatric Illness Utilizing Language and Images Posted to Facebook*, NPJ SCHIZOPHRENIA, Dec. 2020, at 1, 1.

[214] *Id.* at 3.

[215] *Id.* at 1.

[216] *Id.* (noting myriad such studies and lamenting their lack of clinical validity).

[217] For a discussion of the context challenge, see *supra* Section II.A.1.

[218] For a discussion of the classification challenge, see *supra* Section II.A.2.

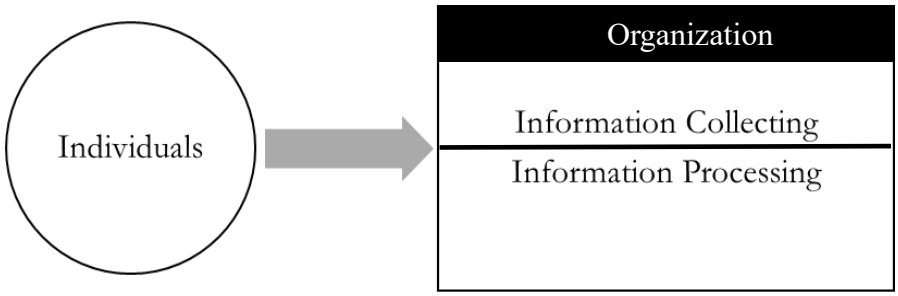
[219] *See* Balkin, *supra* note 30, at 27; *see also* Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. L. REV. CIR. 45, 49 (2015) ("[T]he most important lesson of cyberlaw for robotics is the need to attend to the relationships between affordance and imagination, between tools and relations of power, between technological substrate and social use.").

B. Triangulating Information Privacy in the Inference Economy

This Section argues that most proposed privacy reforms contain an unrecognized structural flaw: they do not appreciate the triangular nature of information privacy dynamics, which ML puts in especially stark relief.

Start with the dominant linear approach to American information privacy protections, as applied to the above examples of bad and good healthcare uses. In the case of the hospital that wanted to assess patient need, patients consented to disclose data to a hospital; that hospital (1) had the data, in its own medical records, and (2) analyzed the data.[220] So, too, with the researchers who assessed knee pain through MRIs; there, patients who consented to the study disclosed information to a single group of researchers who (1) compiled the X-ray data and (2) parsed it to generate a working model.[221] In both of these examples, the collectors were also the processors. In visual terms, the relationship might be depicted roughly as follows, with data flowing from individuals to a single entity that plays two roles:

FIGURE 2



This schematic obscures two points that are essential in the inference economy, in which data about many people can be collected and processed to make inferences about others, and there is an increased potential payoff from engaging in this sort of data analysis. First, it disregards how a particular ML model can be applied back to human beings. The linear approach assumes that the individual who cedes control of their data is the same individual potentially affected by the information collection, processing, or disclosure. In this schema, privacy is personal. But a data-driven ML inference can also be applied to a third party who never entered any agreement.[222]

---

[220] *See supra* notes 204–208 and accompanying text.

[221] *See supra* notes 211–212 and accompanying text.

[222] *See supra* Section II.A.

Second, it fails to underscore that organizations can play distinct roles as data collectors and information processors. Consider a situation like the indeterminate case discussed above: health predictions from internet posts, such as a study that predicts postpartum depression based on social media disclosures.[223] There, the researchers doing the study are the information processors. The original data collector, though, is the social media platform that aggregated the data. A similar division exists in many of the more contentious ML applications. For instance, in a facial recognition instrument, the processing entity might be the same as the collecting entity, as was the case in Facebook's internally created DeepFace model.[224] There, the same pathway as above would still apply. But the entity doing the information processing also might be an unrelated third party, as is the case with, for example, facial recognition company Clearview AI. The linear approach doesn't capture this relational dynamic, either.[225]

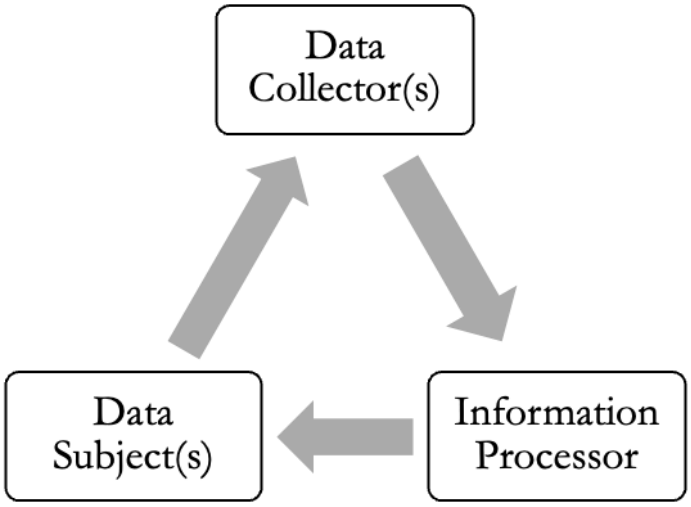
A better approach is to recognize the more complex individual–organizational relationships at stake:

[223] *See* Munmun De Choudhury, Scott Counts, Eric J. Horvitz & Aaron Hoff, *Characterizing and Predicting Postpartum Depression from Shared Facebook Data*, *in* ASS'N FOR COMPUTING MACH., CSCW'14: PROCEEDINGS OF THE 17TH ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK & SOCIAL COMPUTING 625, 626 (2014) (using Facebook data to detect and predict postpartum depression).

[224] *See supra* text accompanying notes 167–168.

[225] *See* Kerr, *supra* note 119, at 132–34 (suggesting that the relationship between the "data subject" and the "other" has become murkier). Rather than focus on the uncertainty of this relationship or the implications for theories of privacy, this Article aims to provide a conceptual framework for thinking about these complex organizational relationships.

FIGURE 3



Here, I use the term "data subject" to cover both (1) an individual whose data is collected, used, or disclosed by an organization or entity and (2) an individual to whom a data-driven ML inference is subsequently applied to derive further information.[226] I use the terms "*data* collector" and "*information* processor" to underscore how the act of processing transforms data to information. It is unnecessary to settle where the "data" versus "information" line falls to denote a phase shift, akin to the change from gas to liquid.[227] Furthermore, by separating "data collector" from "information processor," I do not mean to suggest that these actors are always distinct;

[226] This definition is broader than the one set forth in the GDPR. Under the GDPR, "'personal data' means any information relating to an identified or identifiable natural person ('data subject')," and "an identifiable natural person is one who can be identified, directly or indirectly." 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 (General Data Protection Regulation), art. 4(1), 2016 O.J. (L 119) 1, 3 [hereinafter GDPR]. In a pre-GDPR contribution, Hildebrandt argues that profiling entails a "data subject" in a way that is distinct from conventional data protection legislation. Hildebrandt, *supra* note 9, at 19. Her account similarly suggests that a "subject" can be either the entity from which data is collected or the entity to which an algorithmically generated "profile" is applied. This Article focuses on how these dynamics play out in the American context, with an emphasis on the incentive structures and power dynamics that ML generates.

[227] *See also* Hildebrandt, *supra* note 9, at 17 (positioning data profiling, in general, as a form of knowledge creation). I reserve further consideration of the regulatory consequences of this phase transition, including how it places tremendous pressure on the concept of "personally identifiable information," for future work. Thank you to Nikolas Guggenberger for his incisive questions and comments concerning these categories.

indeed, a company like Google might well occupy both roles. My point is to label the activities as distinct ones.

Like all models, I recognize that this schema simplifies for the sake of expositional clarity. For instance, I do not consider here whether any of the depicted information flows might be bidirectional, and if so, under what conditions.

This representation is nonetheless useful to specify how both the relationships between actors and the dataflows can be different in the ML era.[228] As with the linear approach, data flows between subjects and collectors. It also flows between data collectors and information processors, who aggregate and develop the data into an ML model that is the means to derive more information. Then, the information processor may take the ML working model and apply the prediction to the same person whose data was initially collected. Or it may apply the prediction to other people whom it deems sufficiently similar to a given category of individuals. This cluster of relationships and the power dynamics within it are much more complicated than the linear model.

Even laws that suggest that it is important to take more than two-party relationships into account miss this relational dynamic. European data protection law, for instance, regulates multiple categories of entities that handle individual data and places affirmative obligations on certain entities.[229] Specifically, the European Union's General Data Protection

---

[228] *See* GEORGE E. P. BOX & NORMAN DRAPER, EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACES 74 (1987) ("Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.").

[229] Scholars differ on how much the GDPR's prescriptions create a systemic accountability regime that goes beyond endowing data subjects with individual rights or enhancing individual control. For an argument that the GDPR represents a "binary governance" regime of individual rights and system-wide accountability mechanisms, see Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1582–1615 (2019); see also Meg Leta Jones & Margot E. Kaminski, *An American's Guide to the GDPR*, 98 DENV. L. REV. 93, 116–19 (2021) ("[T]he GDPR consists of two approaches to data protection: a set of individual rights and a set of company obligations."), and Margot E. Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-Layered Explanations*, 11 INT'L DATA PRIV. L. 125, 126–27 (2020). *But see, e.g.*, Pałka, *supra* note 108, at 621–22 (characterizing the EU approach as focused on data protection and individual interests, using "technocratic means of decision-making in place of political ones"); Hartzog, *supra* note 43, at 972–73 (characterizing control as "the archetype for data protection regimes" and consent as "the linchpin" of the GDPR). This Article's focus is the American regime, which, as discussed *supra* Part I, is unabashedly individualistic.

Insofar as data protection in general and the GDPR in particular rely on, at least to some extent, individual control, and ML both undermines individuals' capacity to control their data and unravels "their data" as a coherent category, the pressure that ML puts on American protections extends internationally, too. This issue exists even if the GDPR can also be understood to promote systemic accountability measures.

Regulation (GDPR) identifies "data controllers" and "data processors."[230] Under this framework, a "data controller determines the purposes for which and the means by which personal data is processed."[231] Then, the data processor, which may be a third-party entity or may be the same entity as the data controller, "processes personal data only on behalf of the controller."[232] But this relationship is distinct from the triangular one depicted above. Under the GDPR, a data processor has an explicit contractual, or otherwise legally binding, set of specified duties towards a controller.[233] The relationship between controller and processor is defined by law, and there is a general assumption that the processor is an agent of the controller. And, critically, there is no recognition of the transformation of data to information.

So, too, under the California Consumer Privacy Act of 2018 (CCPA), a leading state effort to enact information privacy protections.[234] There, the statute applies to "businesses" that operate for-profit in California and meet certain size or operational thresholds.[235] It requires those businesses to comply with enumerated consumer-privacy rights.[236] The CCPA also stipulates that businesses that sell or share a consumer's personal information with a third party or "service provider or contractor for a business purpose" must enter into an agreement with that third party, service provider, or contractor.[237] Again, the relationship between controller and processor is defined by law, and there is an assumption that there is a defined, preestablished relationship between a business and another party that might receive or process consumer data. The hitch is that this understanding doesn't account for instances in which there is not such a clear relationship between entities that hold data (data collectors) and entities that develop data into ML models (information processors). Nor does it consider that what is collected may be transformed by the act of processing.

To see the distinct relational dynamics at stake today, consider PimEyes, a publicly accessible website. It bills itself as an "online face search engine" that uses ML technology to allow any individual to "track

---

[230] *See* GDPR, *supra* note 226, art. 4(7)–(8).

[231] *What Is a Data Controller or a Data Processor?*, EUR. COMM'N, https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controller-processor/what-data-controller-or-data-processor_en [https://perma.cc/XUK8-GB6P] (emphasis omitted).

[232] *Id.* (emphasis omitted).

[233] *Id.*

[234] See *supra* text accompanying notes 69–72 (discussing CCPA) and note 69 (discussing subsequent referendum that clarified consumer rights under the statute and created a state privacy protection agency).

[235] CAL. CIV. CODE § 1798.140(d).

[236] *Id.* §§ 1798.100–.125; *California Consumer Privacy Act (CCPA)*, STATE CAL. DEP'T JUST., https://oag.ca.gov/privacy/ccpa [https://perma.cc/9NTT-JS2K].

[237] CAL. CIV. CODE § 1798.100(d).

down [their] face on the Internet, reclaim image rights, and monitor [their] online presence."[238] Individuals are meant to search only for their own image, but there is no binding restriction on who can upload photos to the site or who can receive results.[239] According to media coverage, if provided with a picture of a given individual, PimEyes can scan over 900 million images from across the internet in under one second and find images that match that person.[240] It operates by "crawl[ing]" the web with "bots," "scanning for photos of faces and then recording those images as numerical code."[241] The code is then matched to new images that it scans and receives. In short, this mode of operation works by using ML to take data (photos) and then parsing them against other information (compilations of photos that it has processed into numerical codes) to make inferences about identity.

Tabling specific critiques of PimEyes and the modes of surveillance it facilitates, a close look at the PimEyes model reveals lessons about the broader information privacy ecosystem in the United States. It's possible to understand PimEyes as a "data controller" that also acts as a "data processor" (in GDPR terms) or a "business" (in CCPA terms). But rushing to this conclusion obscures as much as it clarifies. The PimEyes approach relies on the compilations of other data collectors—the masses of data from Google, public sources, and social media sites—to carry out processing.[242] Moreover, there is no relationship—contractual, legal, or implied—between PimEyes and these other entities. Nor are the individuals who agreed to disclose data to these other entities necessarily even aware of PimEyes. In fact, some of those individuals may have done no more than move about in public. And

---

[238] *Face Search Engine Reverse Image Search*, PIMEYES, https://pimeyes.com/en [https://perma.cc/XQW6-PEXJ].

[239] *See More About PimEyes' Database and Opt-Out Service*, PIMEYES, https://pimeyes.com/en/blog/more-about-pimeyes-database-and-opt-out-service [https://perma.cc/L94S-SDQ3]; *see also* Metz, *supra* note 25 (expressing alarm that "PimEyes is open to anyone with internet access").

[240] Drew Harwell, *This Facial Recognition Website Can Turn Anyone into a Cop — or a Stalker*, WASH. POST (May 14, 2021, 7:00 AM), https://www.washingtonpost.com/technology/2021/05/14/pimeyes-facial-recognition-search-secrecy [https://perma.cc/MBJ2-Q47P].

[241] *Id.*

[242] Although the company states that its results come only from publicly accessible sources, researchers have located results that appear to come from social media sites like Instagram, Twitter, YouTube, and TikTok. *Compare Image Search with PimEyes, How to Reverse Image Search*, PIMEYES, https://pimeyes.com/en/blog/image-search-with-pimeyes [https://perma.cc/Y2TL-C997] (stating that PimEyes' search results "come from publicly available websites like the news, media, blogs, company websites, etc." and not from "social media or video platforms, including public profiles (e.g. of companies, influencers, brands)"), *with* Harwell, *supra* note 240 (reporting that "photos from [social media] are regularly among the [PimEyes] results" and there is evidence of search "results from Instagram, YouTube, Twitter and TikTok").

PimEyes is not the only entity that trades on data in the inference economy in this fashion.

Information privacy regulation as we know it misses these relational dynamics. The current regulatory vocabulary does not identify the distinct position that information processors occupy. It thus fails to recognize the distinct power that they can exercise, whether by leveraging data that they themselves collect or by accessing data collected by other entities. This omission is a problem because processing power matters in data analytics in general, and for ML in particular. The activities of information processors, who ingest data from data collectors, can threaten information privacy interests at least as acutely as other entities. And data collectors who are also information processors occupy a particularly powerful inferential position. More clearly labeling the categories "data collector" and "information processor" opens different avenues for reform, targeted at particular legs of the triangle, and helps to more clearly pinpoint which power dynamic a given intervention might address as well as how to do so without foreclosing data analysis that might be socially beneficial.[243]

A triangular framework provides the strategic framework that is missing from leading reform proposals. The remainder of this Section considers interventions at each of the legs of the triangle, highlighting how this reframing casts light on the relationships and dependencies among particular sets of actors, fosters a more nuanced understanding of regulatory objectives, and creates opportunities for novel regulatory interventions to promote information privacy protection at the level of the system, and not merely the level of the individual.[244]

---

[243] Unless otherwise indicated, for ease of exposition, I use the terms "collectors" and "data collectors" and "processors" and "information processors" synonymously in the remainder of this Section. So, too, does the abbreviated word "subject" refer to both senses of the term "data subject." *See supra* text accompanying note 226.

[244] In making these suggestions, I do not advocate an Americanized version of the GDPR. I do think that the U.S. protective regime is missing systemic accountability mechanisms, which some scholars believe the GDPR generates. *See supra* note 229. However, particularly in the American context, where the conditions for GDPR-style "collaborative governance" do not exist, such an approach is misguided. *See* Chander et al., *supra* note 3, at 1761–62 (documenting distinct "legal setting[s]" for the GDPR and CCPA); Hartzog & Richards, *supra* note 53, at 1692 (noting "trans-Atlantic differences in rights, cultures, commitments, and regulatory appetites"); *cf.* Solow-Niederman, *supra* note 144, at 633 (contending that contemporary imbalance of public and private resources, expertise, and power in the United States makes collaborative governance infeasible for AI). I worry, moreover, that a data protection regime will overlook the nature of the relationship among information processors and data collectors and fail to pinpoint relational dependencies that are auspicious intervention points. The present account thus operates one level up and aims to reframe the nature of the relationships at issue in order to clarify the power dynamics and incentives that are salient for subjects, collectors, and processors; catalyze discussion concerning the socially desirable level of data processing in light of those relational dynamics; and, in turn, craft interventions that reflect that determination in a way that is responsive to the American political and legal context. Thank you to Hannah Bloch-Wehba for helpful conversations on this point.

1. Data Subjects and Data Collectors

The subject–collector relationship is most familiar to information privacy law. It seems to fit neatly within the linear approach: data flows between the source of the data (subject) and the entity that aggregates it (collector). Yet concluding that the linear approach amply addresses this relationship is too simplistic. This conclusion does not recognize the ways in which ML exposes and exacerbates latent flaws in the linear, control-centered paradigm. A better strategy is to situate subjects and collectors as one leg of a triangle, with an emphasis on the relationships between the entities, and not the dataflow itself. Doing so highlights the inferential power that collectors may amass relative to subjects.

This framing builds from an emerging scholarly consensus that information privacy is fundamentally and irreducibly relational. An increasing number of privacy scholars focus on privacy as trust, emphasizing the "informational relationships" that define the contemporary era.[245] Complementary work envisions data collectors as entities that, by virtue of how they are granted access to our information, owe particular relational duties to us.[246]

Recognizing that data collectors may or may not be the same entities as information processors allows better tailoring of "information fiduciary" duties and ancillary duties such as a duty of loyalty.[247] These proposals tend to emphasize the relationship between a user and the entity that initially acquires the information as part of a commercial transaction. For instance, in advocating a duty of loyalty for privacy law, Richards and Hartzog argue that dominant regulatory approaches "have overlooked how companies who interact with people in online environments exploit their structural and informational superiority over the people trusting them with their data and

---

[245] *See, e.g.*, Neil Richards & Woodrow Hartzog, *Privacy's Trust Gap: A Review*, 126 YALE L.J. 1180, 1185 (2017) (reviewing FINN BRUNTON & HELEN NISSENBAUM, OBFUSCATION: A USER'S GUIDE FOR PRIVACY AND PROTEST (2015)). For a small sampling of a large body of work that conceptualizes privacy as trust, see generally WALDMAN, *supra* note 7, and Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. REV. 431, 431 (2016).

[246] *See, e.g.*, Balkin, *supra* note 203, at 1185–86 (discussing how "people and organizations who collect and use [personal] data" grow increasingly powerful and should "have special duties to act in ways that do not harm the interests of the people whose information they collect, analyze, use, sell, and distribute"); Jack M. Balkin & Jonathan Zittrain, *A Grand Bargain to Make Tech Companies Trustworthy*, ATLANTIC (Oct. 3, 2016), https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346 [https://perma.cc/4ET3-F76R] (arguing that these data collectors should "have legal obligations to be trustworthy" (emphasis omitted)); *see also* Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961, 966–67 (2021) (developing a duty of loyalty as "a basic element of U.S. data privacy law").

[247] Richards & Hartzog, *supra* note 246, at 995–96 (explaining that the duty of loyalty is usually seen as a primary fiduciary duty and proposing that it could "act as an interpretive guide for other rules and duties").

online experiences."[248] Richards and Hartzog therefore propose that data collectors should owe a duty of loyalty under certain relationally defined conditions,[249] one of which is when "people are made vulnerable" by "their exposure," including through "[t]he collection and processing of personal data."[250]

So, too, is the information fiduciary model centrally concerned with a similar kind of relationship between a company and its users. This model, first articulated by Jack Balkin and later developed along with Jonathan Zittrain, proposes "special duties with respect to personal information that [entities] obtain in the course of their relationships with their [users]."[251] In a world where data and information are synonymous, regulating collectors suffices to protect users who provide data. Moreover, in a world where a collector is also a processor who uses data to generate information, it might well make sense to target the collector as a means to impose corollary duties on the processor. For instance, under a strong form of the information fiduciary model, the collector must ensure that "privacy protections run with the data,"[252] whether subsequent processing occurs inside or outside of the collecting organization.[253]

For data privacy as a whole, more explicitly recognizing the functional roles that data collectors and information processors occupy, and their relationship to one another, permits more precise calibration of the nature and scope of any fiduciary duties owed to subjects. Collectors who are also processors and who have a formal relationship with subjects occupy a

---

[248] *Id.* at 978.

[249] *Id.* at 994 ("We believe that in most circumstances, a duty of loyalty should mean that data collectors are obligated to pursue the 'best interests' of the trusting party with respect to what is exposed and entrusted."); *id.* at 1004 (identifying threshold conditions for the duty of loyalty to apply: "(1) when trust is invited, (2) from people made vulnerable by exposure, (3) when the trustee has control over people's online experiences and data processing, and (4) when people trust data collectors with their exposure"). Earlier work on fiduciary law also focuses on data collectors. *See, e.g.*, Ian Kerr, *Personal Relationships in the Year 2000: Me and My ISP*, *in* PERSONAL RELATIONSHIPS OF DEPENDENCE AND INTERDEPENDENCE IN LAW 78, 84–85 (2002) (discussing how Internet service providers collect data on users and identifying the relationship as one of "dependence").

[250] Richards & Hartzog, *supra* note 246, at 1006.

[251] Balkin, *supra* note 203, at 1208; *see* Jack M. Balkin, *Information Fiduciaries in the Digital Age*, BALKANIZATION (Mar. 5, 2014, 4:50 PM), http://balkin.blogspot.com/2014/03/information-fiduciaries-in-digital-age.html [https://perma.cc/P7LT-QVWB]; *see also* Balkin & Zittrain, *supra* note 246 (advocating for a new kind of law that "clearly states the kinds of duties that online firms owe their end users and customers," including "a duty to look out for the interests of the people whose data businesses regularly harvest and profit from").

[252] Balkin, *supra* note 203, at 1220.

[253] In prior work, I've argued that it makes sense to think of data security in this way, wherein those who obtain data under conditions of trust are held responsible if their choices enable data breaches that violate that trust. Solow-Niederman, *supra* note 56, at 625–26.

particular position of inferential power relative to subjects. Most digital-platform firms, such as Facebook or Google, fall into this category. That's why it may be most appropriate to impose the full information fiduciary framework on such data collectors.[254]

This recognition, moreover, underscores how and why different approaches are required to regulate collectors who are also processors, as compared to other processors. For joint collector-processors, interventions such as bans of a particular kind of ML instrument establish that a collector may not process data in a particular way, on the grounds that doing so inappropriately leverages the firm's position in the inference economy. But processors that are not directly related either to the collector or to the subject—picture an outside firm that gleans "emergent medical data" about mood disorders from the colors in Instagram posts[255]—may require a complementary regulatory approach targeted at that leg of the triangle. Labeling this subset of processors as "collectors" obscures how such processors take advantage of the affordances of other collectors and too easily allows those other collectors to evade responsibility for the broad reach of their choices.[256] A triangular framing of the interests at stake facilitates a more informed policy conversation by highlighting where a trust-based approach may work best, indicating where policymakers or courts may need to be more precise with the terms that they use to refer to the data collector (or information processor) at issue, and suggesting where other interventions to regulate processing activities directly may be required.

2. Data Collectors and Information Processors

The collector–processor relationship represents an underexplored avenue for intervention. The linear approach assumes a direct relationship between collectors and processors. As we have seen, however, that direct relationship is not always present. Processors can draw inferences from compilations of information that are made available by data collectors, whether or not they have any formal relationship to those collectors. And ML opens up precisely these inferential pathways.

A trilateral relational frame highlights how collectors' choices make data more or less available, and how these choices in turn affect what activities processors may execute. Put differently, collectors' decisions determine how easy or hard it is to compile information. This compilation

---

[254] *Cf.* Richards & Hartzog, *supra* note 53, at 1746 (noting "stringent duties" of the information fiduciary model and calling for a complementary set of "trust rules" that "are not necessarily dependent upon formal relationships to function").

[255] *See* Birnbaum et al., *supra* note 213, at 1; *see also supra* note 106 and accompanying text (discussing "emergent medical data").

[256] *See infra* Section IV.B.2.

matters. Privacy scholars have long warned of the harms that can be unleashed in a world where there are masses of compiled data about individuals.[257] Indeed, this concern with data aggregation and the profits to be reaped from it animates surveillance capitalist critiques;[258] moreover, the 1973 HEW report on privacy was motivated by a concern with the emergence of centralized, computerized databases.[259]

What is new is how processors can now centralize data by compiling aggregated bodies of data that other collectors fail to amply protect and then use this data to derive further information. For instance, although social media posts that mention a sensitive medical condition are not centrally collected by the social media platform, these posts can be understood as distributed data points that are ripe for processing by external actors. How hard or easy a collector makes it to harvest these data points, and with what consequences, affects a processor's access to data in ways that, in turn, limit or expand the kinds of activities that the processor can undertake.

To make this point more concrete, take the example of face datasets and the generation of commercial facial recognition tools. A company like Clearview AI relied on Facebook and other images collected by platforms to generate its database.[260] In the face of mounting public opposition to facial recognition databases, including several mainstream-media exposés, Facebook went on the record to chastise Clearview AI.[261] Other companies such as Twitter, YouTube, and Venmo have also publicly stated that Clearview's scraping practices violate their terms of service.[262] These firms seem to have limited their responses to cease-and-desist letters and public denunciations, after the scraping was already done (and only in the wake of mounting public controversy about facial recognition technologies).

---

[257] *See* Ohm, *supra* note 52, at 1746 (describing "database[s] of ruin," or the potential for "the worldwide collection of all of the facts held by third parties that can be used to cause privacy-related harm to almost every member of society"); Danielle Keats Citron, *Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age*, 80 S. CAL. L. REV. 241, 244 (2007) (arguing that computer databases containing personal, identifying information should be understood as "reservoirs" that endanger the public if they leak).

[258] *See* Mariano-Florentino Cuéllar & Aziz Z. Huq, *Economies of Surveillance*, 133 HARV. L. REV. 1280, 1295–97 (2020) (reviewing ZUBOFF, *supra* note 16).

[259] ADVISORY COMM. ON AUTOMATED PERS. DATA SYS. , *supra* note 47, v–vi.

[260] *See* Hill, *supra* note 4.

[261] *See* Steven Melendez, *Facebook Orders Creepy AI Firm to Stop Scraping Your Instagram Photos*, FAST CO. (Feb. 6, 2020), https://www.fastcompany.com/90461077/facebook-joins-fellow-tech-companies-in-publicly-opposing-a-controversial-face-recognition-firm [https://perma.cc/7Pj9-AMM6].

[262] *Facebook, Twitter, Youtube, Venmo Demand AI Startup Must Stop Scraping Faces from Sites*, ASSOCIATED PRESS (Feb. 5, 2020, 10:16 PM), https://www.marketwatch.com/story/facebook-twitter-youtube-venmo-demand-ai-startup-must-stop-scraping-faces-from-sites-2020-02-05 [https://perma.cc/K4NL-XDBE].

These companies could have done more, and sooner. For instance, on the technological side, such firms could have implemented an automated flag whenever an entity scraped a suspiciously large quantity of data from the site, creating an early warning system before an entity like Clearview processed the data. And on the legal side, these firms could have stepped up enforcement of their terms of service with litigation. The choice neither to implement technical measures nor to advocate on behalf of their users' interests in the court of public opinion or in actual court was an active decision by collectors.[263] And that decision facilitated processing by parties with no relationship to the collectors' users.[264] A triangular framing

[263] That's not to say that lawsuits would have been a slam dunk, particularly if brought under the Computer Fraud and Abuse Act (CFAA). Some of these scraping activities occurred in the shadow of a 2019 CFAA case, *hiQ Labs, Inc. v. LinkedIn Corp.*, which involved a dispute between LinkedIn and a rival corporate-analytics company that had scraped information posted on public-facing portions of LinkedIn profiles. 938 F.3d 985, 989–92 (9th Cir. 2019). The Ninth Circuit found "serious questions about whether LinkedIn may invoke the CFAA to preempt hiQ's possibly meritorious [state law] tortious interference claim." *Id.* at 1004. In June 2021, the Supreme Court granted LinkedIn's petition for a writ of certiorari, vacated the Ninth Circuit's judgment, and remanded the case "for further consideration in light of" the Court's disposition of a different CFAA suit, *Van Buren v. United States*, 141 S. Ct. 1648 (2021), which narrowed the statute's reach. *See* LinkedIn Corp. v. hiQ Labs, Inc., 141 S. Ct. 2752 (2021); Orin S. Kerr, *Focusing the CFAA in* Van Buren, 2021 SUP. CT. REV. 155, 164–65, 173–80 (explaining how *Van Buren* "partially focuses" the CFAA picture); Orin Kerr, *The Supreme Court Reins in the CFAA in* Van Buren, LAWFARE (June 9, 2021, 9:04 PM), https://www.lawfareblog.com/supreme-court-reins-cfaa-van-buren [https://perma.cc/5ND9-2W7L].

In April 2022, hearing the case on remand, the Ninth Circuit arrived at the same conclusion that it had reached previously and affirmed the district court's initial disposition granting hiQ's motion for a preliminary injunction, finding that public profiles do not require authorization or access permission and thus are not subject to the access limitations set forth in the CFAA. hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1202 (9th Cir. 2022) (on remand from the U.S. Supreme Court). The Ninth Circuit distinguished between the publicly available information at issue in *hiQ Labs* and situations in which a website "'has tried to limit and control access to its website' as to the purposes for which . . . [an outside entity] sought to use it." *Id.* at 1199 (quoting Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016)). It also left open, without deciding, the possibility that "web scraping exceeding the scope of the website owner's consent gives rise to a common law tort claim for trespass to chattels, at least when it causes demonstrable harm." *Id.* at 1202 n.21.

[264] *See* Jonathan Zittrain & John Bowers, *A Start-Up Is Using Photos to ID You. Big Tech Can Stop It from Happening Again.*, WASH. POST (Apr. 14, 2020, 3:58 PM), https://www.washingtonpost.com/outlook/2020/04/14/tech-start-up-is-using-photos-id-you-big-tech-could-have-stopped-them [https://perma.cc/439Y-JBCQ] (suggesting "platforms must shoulder some of the blame" for Clearview AI's development). I do not mean to suggest that enforcement under a statute like the CFAA is necessarily a good idea, at least without substantial clarification of the statute. For instance, it seems important, as a policy matter, to distinguish between access for research and access for commercial purposes. *See* SUNOO PARK & KENDRA ALBERT, HARV. L. SCH. CYBERLAW CLINIC & ELEC. FRONTIER FOUND., A RESEARCHER'S GUIDE TO SOME LEGAL RISKS OF SECURITY RESEARCH 8 (2020). It is essential to think carefully about how to draw the right lines between access to publicly accessible information and access to information that the user of a platform service believes is private. For an argument that the use of cyber-trespass laws like the CFAA to bar access to publicly available information amounts to a First Amendment violation, see Thomas E. Kadri, *Platforms as Blackacres*, 68 UCLA L. REV. 1184, 1190–93 (2022).

underscores not only this facilitation, but also processors' dependency on collectors.

Furthermore, a triangular approach reveals how the regulatory status quo, coupled with the business model of platform firms, incentivize arrangements that align collectors and processors against subjects' interests. For example, media reports allege that Clearview scraped profile images from the payment platform Venmo.[265] Venmo exposed any profile photos that a user has ever uploaded, simply by manually changing the image URL, and did not provide any direct way for Venmo users to delete or even to review these images.[266] The work of the processor (Clearview) is possible in no small part because of the choices of the collector (Venmo). At present, the informational power that flows from that relationship is essentially unchecked, apart from companies' own choices.

Excavating these relational dependencies reveals intervention points that emphasize the collector–processor leg of the triangle. For instance, on the regulatory side, the FTC could undertake a set of strategic enforcement activities against firms that do not enforce their own terms of service against third-party violators.[267] Alternatively, or in addition, a body within the FTC, such as the new rulemaking group proposed by former Acting FTC Chair Rebecca Slaughter, could issue a statement concerning this third-party evasion of firms' terms of service, thereby providing a roadmap for collectors to follow.[268] These rules would need to provide more than thin procedural guidance and would need to avoid conflating consumer consent with meaningful control over actual information flows. They would need to specify the minimum standard that platforms that collect data must follow when enforcing their own terms of service, thereby creating a floor below

---

[265] *See* Hill, *supra* note 4; *Facebook, Twitter, YouTube, Venmo Demand AI Startup Must Stop Scraping Faces from Sites*, *supra* note 262; Louise Matsakis, *Scraping the Web Is a Powerful Tool. Clearview AI Abused It*, WIRED (Jan. 25, 2020, 7:00 AM), https://www.wired.com/story/clearview-ai-scraping-web [https://perma.cc/6T5D-W655]; Ryan Mac, Caroline Haskins & Logan McDonald, *Clearview AI Has Promised to Cancel All Relationships with Private Companies*, BUZZFEED NEWS (May 7, 2020, 5:50 PM), https://www.buzzfeednews.com/article/ryanmac/clearview-ai-no-facial-recognition-private-companies [https://perma.cc/WWE4-VSDT].

[266] *See* Katie Notopoulos, *Venmo Exposes All the Old Profile Photos You Thought Were Gone*, BUZZFEED NEWS (May 18, 2021, 8:29 AM) https://www.buzzfeednews.com/article/katienotopoulos/paypals-venmo-exposes-old-photos?mc_cid=da82a8d945&mc_eid=0cfb8ad92b [https://perma.cc/7JUZ-P8SN].

[267] There is administrative law precedent for this move. *See* Solove & Hartzog, *supra* note 56, at 663 (citing FTC v. Accusearch Inc., No. 06-CV-0105 (D. Wyo. Sept. 28, 2007)) (discussing a 2007 FTC enforcement action in which the Commission asserted that one company engaged in unfair practices by facilitating another company's violation of the Telecommunications Act).

[268] *See* Press Release, Fed. Trade Comm'n, FTC Acting Chairwoman Slaughter Announces New Rulemaking Group (Mar. 25, 2021), https://www.ftc.gov/news-events/press-releases/2021/03/ftc-acting-chairwoman-slaughter-announces-new-rulemaking-group [https://perma.cc/PS8G-6VEP].

which acceptable business practices should not fall. Guidance of this sort would not only help users, but also provide a more predictable environment for firms by clarifying what is expected of them with respect to external processors.

Such administrative guidance might be most effective if paired with technical solutions to help regulated collectors comply with any such formal guidance. Technical interventions might automatically identify widespread scraping of a website. Specifically, because so-called "bots" that scrape websites tend to operate at far faster speeds than human users, websites might monitor the speed of interactions with the site to create a signal that scraping is likely occurring.[269] The FTC or other regulatory bodies might then explicitly incorporate technical interventions of this sort into published guidance on "Privacy by Design";[270] over time, these standards could become part of the expected set of standard privacy practices for firms that trade in data. In addition, as the next Section addresses, a more explicit focus on the subject–processor dynamic facilitates a more textured understanding of subjects' interests relative to each of these parties.

3. Data Subjects and Information Processors

A triangular frame directs attention to subject–processor relationships that the linear model tends to obscure. Processing is relevant within traditional frames—but primarily in terms of dataflows. For example, data protection regulations can and do consider use restrictions,[271] and the FTC's unfair and deceptive trade practices analysis may take into account whether individuals agreed to the full suite of processing activities at the time they consented to terms of service.[272]

Homing in on the subject–processor leg of the triangle forces greater specificity about why a particular information-processing activity, as applied to data subjects, warrants attention. Two categories of issues stand out. One concerns the processor: what kinds of processors are positioned to leverage data in the inference economy, subject to what constraints? The other

---

[269] *See What Is Data Scraping?*, CLOUDFLARE, https://www.cloudflare.com/learning/bots/what-is-data-scraping [https://perma.cc/S5V3-2FJF].

[270] In 2012, the FTC adopted privacy by design as a baseline principle in its privacy-framework report. *See* FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESS AND POLICYMAKERS 22–32 (2012).

[271] The GDPR, for instance, includes "purpose limitation[s]" on even lawfully collected data, and the proposed implementing regulations for the CCPA, which were not included in the final text, stipulated that a business "shall not use a consumer's personal information for a purpose materially different than those disclosed in the notice at collection." *See* Chander et al., *supra* note 3, at 1756–57 (quoting CAL. CODE REGS. tit. 11, § 999.305(a)(5) (withdrawn July 29, 2020)).

[272] *See supra* notes 86–91 and accompanying text (discussing the EverAlbum settlement).

concerns the subject: what is the felt impact of a processing decision at the point of application by an ML instrument?

Take, first, the processor. Because what a processor can do turns on both formal law and how technological and economic configurations constrain or enable particular activities, it would be incorrect to suggest that any old processor can vacuum up publicly available data and efficiently convert it into a working algorithm. To the contrary, the current state of ML generally requires vast amounts of data and compute to operate effectively.[273] Thus, the areas in which widespread ML analytics are possible will depend at least in part on access to these resources.

Two processor-related insights follow. One, because access to adequate compute tends to be concentrated in a comparatively small set of firms rather than democratized, mergers and acquisitions have profound implications for data privacy. Indeed, contemporary privacy regulators might do well to take a page from competition law and consider how the accumulation of hardware and data capital can erode structural protections of privacy by facilitating a wider range of processing activities.[274] As a case in point, when the FTC approved Google's acquisition of the online ad-serving company DoubleClick, it provided Google with a vast new reservoir of data to process—despite the objections of privacy advocates who were concerned that the merger was not in the public interest for this very reason.[275] Future policymakers should pay close attention to similar privacy risks not only for data acquisition, but also for further concentrations of compute power.

Two, the available range of processing activities is contingent on the technological state of the art. Further changes in compute power, such as a major breakthrough in quantum computing, could significantly alter the political economy of information privacy.[276] So, too, could rapid progress on ML models that permit efficient training with less data,[277] or the realization

---

[273] *See supra* Section II.B.

[274] In a future project, *(De)Platforming Artificial Intelligence*, I intend to address the political economy of AI development in more detail.

[275] *See* Dawn Kawamoto, *FTC Allows Google-DoubleClick Merger to Proceed*, CNET **(**Mar. 21, 2008, 1:52 PM**),** https://www.cnet.com/news/ftc-allows-google-doubleclick-merger-to-proceed-1 [https://perma.cc/43Y6-LDC6].

[276] *See* Lohr, *supra* note 159 (discussing how the need for compute power leads to centralization in AI). For further analysis of how computational power shapes AI development paths, see Tim Hwang, Computational Power and the Social Impact of Artificial Intelligence (Mar. 23, 2018) (unpublished manuscript), https://papers.ssrn.com/a=3147971 [https://perma.cc/ZY3V-7ZWX].

[277] These methods are alluring given their transformative potential yet remain largely theoretical. *See* Karen Hao, *A Radical New Technique Lets AI Learn with Practically No Data*, MIT TECH. REV. (Oct. 16, 2020), https://www.technologyreview.com/2020/10/16/1010566/ai-machine-learning-with-tiny-data [https://perma.cc/964C-JQ3W] (discussing efforts to create "less than one shot" learning capable of

of a National Research Cloud that increases the supply of data to trusted actors.[278] Focusing on processors as distinct entities brings these considerations into the frame of information privacy regulation.

Furthermore, an emphasis on the subject–processor relationship directs attention to the people affected by a particular data-driven model. For instance, in thinking about information processing, there is a meaningful distinction between a tool that has a discriminatory effect on individuals, even if it is developed and trained with representative data, and a tool that has the potential for discriminatory impacts if it is trained on a nondiverse dataset or otherwise does not follow best practices in its development. The first example—a processing activity that has a high risk of biased informational outputs, no matter what—presents the strongest justification for a ban. Emotion-recognition technologies, which inevitably require blunt racial and cultural judgments about how individuals' faces look when they present certain emotions, might fall into this category.[279] Any woman who has been accused of having "resting bitch face" when she is merely thinking knows the problem all too well.[280] In such situations, bright-line rules may be most appropriate.

"recogniz[ing] *more* objects than the number of examples it was trained on"); Natalie Ram, *One Shot Learning in AI Innovation*, AI PULSE (Jan. 25, 2019), https://aipulse.org/one-shot-learning-in-ai-innovation/?pdf=142 [https://perma.cc/ALP5-252E] (discussing developing efforts to create one-shot learning models that can be trained with less data).

[278] *See* John Etchemendy & Fei-Fei Li, *National Research Cloud: Ensuring the Continuation of American Innovation*, STAN. UNIV. HUM.-CENTERED A.I. (Mar. 28, 2020), https://hai.stanford.edu/news/national-research-cloud-ensuring-continuation-american-innovation [https://perma.cc/DJV9-SS6P] (advocating for a National Research Cloud to provide data and compute for academic researchers). The National Artificial Intelligence Initiative Act of 2020 directs the National Science Foundation Director and the Office of Science and Technology Policy to "investigate the feasibility and advisability of establishing and sustaining a National Artificial Intelligence Research Resource," which would include shared compute power and access to government datasets. *See* National Artificial Intelligence Initiative Act of 2020, Pub. L. No. 116-283, § 5106, 134 Stat. 4523, 4531–34 (codified as amended at 15 U.S.C. § 9415), https://www.congress.gov/116/plaws/publ283/PLAW-116publ283.pdf [https://perma.cc/V9EZ-LDMW].

[279] *See, e.g.*, Luke Stark, *The Emotive Politics of Digital Mood Tracking*, 22 NEW MEDIA & SOC'Y 2039, 2040–41 (2020), https://journals.sagepub.com/doi/abs/10.1177/1461444820924624 [https://perma.cc/BK8W-KF4C] (assessing the impact of cultural influences in the context of digital mood-tracking applications Moodscope and MoodPanda); *see also* Mark Purdy, John Zealley & Omaro Maseli, *The Risks of Using AI to Interpret Human Emotions*, HARV. BUS. REV. (2019), https://hbr.org/2019/11/the-risks-of-using-ai-to-interpret-human-emotions [https://perma.cc/M77E-UUGT] (considering the risks of bias in emotional AI technology and noting how "one study found that emotional analysis technology assigns more negative emotions to people of certain ethnicities than to others").

[280] *See* Jessica Bennett, *I'm Not Mad. That's Just My RBF.*, N.Y. TIMES (Aug. 1, 2015), https://www.nytimes.com/2015/08/02/fashion/im-not-mad-thats-just-my-resting-b-face.html [https://perma.cc/T3AL-J673].

The second example—a processing activity that is problematic because of flawed implementation—might call for standards that guide development choices and thereby regulate how a processor can affect subjects. Congress would not need to legislate to generate such standards; there are several regulatory avenues available. For one, the FTC could consider providing more substantive guidance concerning what it means for a dataset to be adequately diverse through rulemaking, notwithstanding the procedural burdens to which it is subject.[281]

Alternatively, or additionally, agencies responsible for regulating processing in especially sensitive domains could revisit the specificity of the regulatory guidance that they provide. As one example, consider lending laws. The FTC has emphasized that "[t]he lending laws encourage the use of AI tools that are 'empirically derived, demonstrably and statistically sound.'"[282] This informal guidance references Regulation B, promulgated by the Consumer Financial Protection Bureau (CFPB). Regulation B provides that a tool that is "demonstrably and statistically sound" must be "[d]eveloped and validated using accepted statistical principles and methodology."[283] But this procedural guidance only goes so far when it comes to AI-powered tools. Fairness in ML is hotly contested.[284] There are no "accepted statistical principles and methodology" in many ML contexts; rather, the very choice of a mathematical definition of "fairness" is a political one.[285]

---

[281] The FTC lacks general rulemaking authority under the Administrative Procedure Act (APA) or specific authority to issue information privacy rules. *See* COHEN, *supra* note 17, at 188 (discussing the "FTC's practice of lawmaking through adjudication"). The contemporary Commission instead has Magnuson-Moss (Mag-Moss) rulemaking authority. *See* Magnuson-Moss Warranty—Federal Trade Commission Improvement Act, Pub. L. No. 93-637, 88 Stat. 2183 (1975) (codified as amended at 15 U.S.C. §§ 45–46, 49–52, 56–57c, 2301–2312 (2012)). Mag-Moss rulemaking is more procedurally burdensome than APA informal rulemaking procedures. Rebecca Kelly Slaughter, Comm'r, Fed. Trade Comm'n, Remarks at New York University School of Law Cybersecurity and Data Privacy Conference Program on Corporate Compliance and Enforcement: FTC Data Privacy Enforcement: A Time of Change 5–6 (Oct. 16, 2020), https://www.ftc.gov/system/files/documents/public_statements/1581786/slaughter_-_remarks_on_ftc_data_privacy_enforcement_-_a_time_of_change.pdf [https://perma.cc/F2X4-NCAS]. As Julie Cohen notes, because of the limits of its regulatory authority, "the FTC's enforcement posture reflects an especially complex calculus." COHEN, *supra* note 17, at 188.

[282] Andrew Smith, *Using Artificial Intelligence and Algorithms*, FED. TRADE COMM'N (Apr. 8, 2020), https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms [https://perma.cc/8HTS-F3EC] (quoting 12 C.F.R. § 1002.2 (2018) (Regulation B)).

[283] 12 C.F.R. § 1002.2(p) (2018) (Regulation B).

[284] *See* Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli & Richmond Y. Wong, *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, PROC. ACM ON HUM.-COMPUT. INTERACTION, Nov. 2019, at 1, 4–5, 16, https://arxiv.org/pdf/1909.11869.pdf [https://perma.cc/SM8B-STB7].

[285] *See* Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk [https://perma.cc/Z3XX-N8QM].

Attention to the subject–processor leg of the triangle underscores the human beings affected by the act of information processing and foregrounds why process alone cannot answer the substantive question of what is "unfair" here.[286] Technical and social understandings of fairness are not necessarily aligned,[287] and seemingly technical choices such as where to set a threshold in an ML training model can result in outcomes that satisfy a given measure of fairness for some populations but not for others.[288] Furthermore, decisions such as the level of false-positive or false-negative error rate to tolerate are themselves normatively laden.[289] Accordingly, an agency like the CFPB may need to revisit language such as Regulation B to recognize the fact that there may be no settled statistical consensus around, for instance, an acceptable error rate in a tool, or whether false positives or false negatives are more problematic in a given context. That's not to say that the government would be more accurate, however accuracy is measured, than a private firm with a profit motive to be accurate; rather, it's to argue that, in instances that present a high risk of invidiously discriminatory impact, some form of public standard-setting is wise.

To that end, the Commerce Department's National Institute of Standards and Technology (NIST) represents an untapped source of guidance. Specifically, the 2021 National Defense Authorization Act (NDAA) grants NIST the authority to "support the development of technical standards and guidelines" to "promote trustworthy artificial intelligence systems" and "test for bias."[290] NIST is further tasked with developing "a voluntary risk management framework" for AI systems, including "standards, guidelines, best practices, methodologies, procedures and processes" for "trustworthy" systems as well as "common definitions and

---

[286] *See* COHEN, *supra* note 17, at 179–80 (discussing CFPB Regulation B and highlighting how it "leaves unexplained what . . . [the referenced] principles and methods might be and how they ought to translate into contexts involving automated, predictive algorithms with artificial intelligence or machine learning components").

[287] Mulligan et al., *supra* note 284, at 5–6.

[288] *See* Alicia Solow-Niederman, YooJung Choi & Guy Van den Broeck, *The Institutional Life of Algorithmic Risk Assessment*, 34 BERKELEY TECH. L.J. 705, 734–39 (2019); *see also* Rohit Chopra, Comm'r, Fed. Trade Comm'n, Remarks at Asia Pacific Privacy Authorities 54th APPA Forum 2–3 (Dec. 7, 2020), https://www.ftc.gov/system/files/documents/public_statements/1585034/chopra-asia-pacific.pdf [https://perma.cc/UUN5-CEMP].

[289] This issue is by no means academic; to the contrary, recent controversies concerning the use of automated risk-assessment tools have centered on competing understandings of whether a tool can be considered fair when it has different false-positive and false-negative error rates for different demographic groups. *See, e.g.*, Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/F3XJ-DQ98].

[290] William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5301 (2021).

characterizations for aspects of trustworthiness, including explainability, transparency, safety, privacy, security, robustness, fairness, bias, ethics, validation, verification, interpretability, and other properties related to artificial intelligence systems that are common across all sectors."[291] Whether or not NIST can achieve this ambitious target will likely depend in part on how much the agency hews to a strictly "technical" as opposed to a more socially-informed understanding of standard-setting.[292] And voluntary standards are no panacea, particularly given the outsized private influence in the ML industry.[293]

By delineating the minimum technical rules of the road, such standards can nonetheless usefully set floors for acceptable information processing. These floors can in turn provide regulatory hooks for agencies that monitor the limits of processing and suggest common law standards for courts that encounter any tort or contract law claims about ML processing.[294] The subject–processor framing draws attention to the manner in which these kinds of technical standards can affect ML's development path, thereby regulating how ML models affect people on the ground.

* * *

The inference economy is a reality. We cannot account for it if we are insufficiently attentive to the ways in which informational power is distributed among data subjects, data collectors, and information processors. These dynamics are meaningfully distinct from those assumed in conventional privacy regulations. Triangulating information privacy as the result of these relationships both provides a strategic framework that is better calibrated for institutional power dynamics and opens pathways to more effective tactical interventions.

---

[291] *Id.*; *see also Summary of AI Provisions from the National Defense Authorization Act 2021*, STAN. UNIV. INST. HUM.-CENTERED A.I., https://hai.stanford.edu/policy/policy-resources/summary-ai-provisions-national-defense-authorization-act-2021 [https://perma.cc/WNH4-M3XQ] (discussing Section 5301 of the 2021 National Defense Authorization Act).

[292] *See* Solow-Niederman, *supra* note 144, at 693 (2020) (arguing that "public actors can and should place a greater emphasis on the 'non-technical' standards . . . that 'inform policy and human decision-making.'" (quoting NAT'L INST. OF STANDARDS & TECH., U.S. LEADERSHIP IN AI: A PLAN FOR FEDERAL ENGAGEMENT IN DEVELOPING TECHNICAL STANDARDS AND RELATED TOOLS 13 (2019))).

[293] *Id.* at 675–80 (describing the resource imbalances between public and private players in the context of AI development).

[294] *See* Frank Pasquale, *Data-Informed Duties in AI Development*, 119 COLUM. L. REV. 1917, 1920 (2019).

CONCLUSION

The inference economy challenges information privacy. That's because information privacy protections rely on linear, control-centered frameworks that ask for individual consent and then open or close dataflows based on that consent. But information flows do not start and end with one person's control over their personal data. Seemingly innocuous or irrelevant data can generate ML insights, making it impossible for an individual to predict what kinds of data are important to protect. Moreover, it is possible to aggregate myriad individuals' data within ML models, identify patterns, and then apply those patterns to make probabilistic inferences about other people. As a result, what matters today is not just one individual's control over their personal, identifiable information. It's not even clear that the category "their personal, identifiable information" is the right one on which to focus in a world where aggregated data that is neither personal nor identifiable can be used to make inferences about you, me, and others. Our world features an altogether different epistemic pathway from data to information to knowledge.

The contemporary reality is an inference economy. The inference economy consists of a network of relationships to manage—not a set of dataflows for individuals to constrain. Preserving information privacy protection today requires recognizing a historically overlooked relationship that machine learning makes particularly salient: the connections between those who access and amass data and those who subsequently process it to draw inferences. Rather than double down on unresponsive, control-centered tactics, a better strategy is to focus on the relationships that emerge in a more complex, triangular model of data subjects, data collectors, and information processors, and to develop regulatory interventions with an eye to who has amassed informational power at each leg of the triangle, how they have done so, and to what effect they use the data. Privacy protection in the inference economy requires confronting which organizations have capabilities and incentives to do things with data. We ignore that at our peril.