DATA AND DEMOCRACY



Licensure as Data Governance

By Frank Pasquale

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In October 2020, the Knight First Amendment Institute at Columbia University convened a virtual symposium, titled "Data and Democracy," to investigate how technological advances relating to the collection, analysis, and manipulation of data are affecting democratic processes, and how the law must adapt to ensure the conditions for self-government. This symposium was organized by the Institute's 2019-2020 Senior Visiting Research Scholar, Yale Law Professor Amy Kapczynski, and co-sponsored by the Law and Political Economy Project at Yale Law School.

The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, information studies, political science, and other disciplines, the essays focus on three areas that are both central to democratic governance and directly affected by advancing technologies and ever-increasing data collection: 1) public opinion formation and access to information; 2) the formation and exercise of public power; and 3) the political economy of data.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Amy Kapczynski, Senior Visiting Research Scholar; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; A. Adam Glenn, Writer/Editor; and Madeline Wood, Research Coordinator.

The full series is available at knightcolumbia.org/research/

[In the late 1950s, the U.S.] government abdicated … responsibility to establish rules, safeguards, and standards relating to the collection and use of personal data for the purpose of directing human behavior. Plainly, all of this might have gone differently. Plenty of people believed at the time that a people machine was entirely and utterly amoral. "My own opinion is that such a thing (a) cannot work, (b) is immoral, (c) should be declared illegal," [soon-to-be-FCC Chair] Newton Minow had written to Arthur Schlesinger in 1959. "Please advise."

—Jill Lepore, *If Then: How the Simulmatics Corporation Invented the Future*, 323.

INTRODUCTION

DATA PROTECTORS AND PRIVACY REGULATORS face a crisis of overwork and underresourcing. Enforcement of privacy laws is too often belated, if it comes at all. Massive firms with myriad data points on tens of millions of people face fines for data misuse and security breaches that are the economic equivalent of a parking ticket. Potentially worse than all these well-recognized barriers to accountability is a known unknown: namely, the black box problem. Even the most diligent regulators and civil society groups have little idea of the full scope and intensity of data extraction, analysis, and use at leading firms, given the triple barriers of trade secrecy, nondisclosure agreements, and technical complexity now effectively hiding their actions from public scrutiny. This crisis is likely to continue unless there is a fundamental shift in the way we regulate the collection, analysis, transfer, and use of data.[1]

At present, policymakers tend to presume that the data practices of firms are legal, and only investigate and regulate when there is suspicion of wrongdoing. What if the presumption were flipped? That is, what if a firm had to certify that its data practices met clear requirements for security, nondiscrimination, accuracy, appropriateness, and correctability, before it

collected, analyzed, or used data?[2] Such a standard may not seem administrable now, given the widespread and rapid use of data—and the artificial intelligence (AI) it powers—at firms of all sizes. But such requirements could be applied, at first, to the largest firms' most troubling data practices, and only gradually (if at all) to smaller ones and less menacing data practices. For example, would it really be troubling to require firms to demonstrate basic security practices once they have accumulated sensitive data on over 1 million people, before they continue to collect even more? Scholars have argued that certain data practices should not be permitted at all.[3] Rather than expecting underfunded, understaffed regulators to overcome the monumental administrative and black box problems mentioned above, responsibility could be built into the structure of data-driven industries via licensure schemes that require certain standards to be met before large-scale data practices expand even further.[4]

To give a concrete example motivating this flipped presumption about data practices, consider the emergence of health inferences from data that is not, on its face, health-predictive. For instance, an AI program, reviewing only writing samples, "predicted, with 75 percent accuracy, who would get Alzheimer's disease."[5] This type of inference could be used in subtle or secretive ways by the firms making it, as well as by employers, marketers, financial institutions, and other important decision-makers.[6] Such predictions may have massive impacts on those projected to have Alzheimer's, including denial of life insurance or long-term care insurance, denial of employment, or loss of other opportunities. Even where such uses of the data are illegal, complex and expensive legal systems may make it very difficult to enforce one's rights. Governments should ensure ex ante that predictions are done and used in a responsible way, much as federally funded research is often channeled through institutional review boards in order to respect ethical and legal standards.[7]

A licensure regime for data and the AI it powers would enable citizens to democratically shape data's scope and proper use, rather than resigning ourselves to being increasingly influenced and shaped by forces beyond our control. To ground the case for more ex ante regulation, Part I describes the expanding scope of data collection, analysis, and use, and the threats that that scope poses to data subjects. Part II critiques consent-based models

of data protection, while Part III examines the substantive foundation of licensure models. Part IV addresses a key challenge to my approach: the free expression concerns raised by the licensure of large-scale personal data collection, analysis, and use. Part V concludes with reflections on the opportunities created by data licensure frameworks and potential limitations upon them.

I. THE EXPANDING SCOPE OF DATA COLLECTION, ANALYSIS, AND USE

As data collection becomes more prevalent, massive firms are privy to exceptionally comprehensive and intimate details about individuals.[8] These can include transport, financial, retail, health, leisure, entertainment, location, and many other kinds of data. Once large enough stores of data are created, there are increasing opportunities to create inferences about persons based on extrapolations from both humanly recognizable and ad hoc, machine learning-recognizable groups.[9]

Much observation can occur without persons' consent. Even when consent is obtained, improper data collection may occur. Increasingly desperate individuals may be effectively coerced via their circumstances to permit comprehensive, 360-degree surveillance of key aspects of their lives. For example, many people now cannot access credit because of a "thin credit file"—industry lingo for someone without much of a repayment history. Fintech firms promise "financial inclusion" for those willing to give lenders access to their social media activity and other intimate data.[10] Critics characterize this "opportunity" as predatory inclusion, and it is easy to see why: Those enticed into the sphere of payday lenders may not only contract into unsustainable debt-repayment schemes but may also become marks for other exploitative businesses, such as for-profit universities or unlicensed rehab centers.

The economic logic here (giving up more privacy in exchange for lower interest rates or other favorable terms) is, in principle, illimitable. For example, a firm may give a loan applicant a reduced rate on a car loan, if she allows it (and its agents) to download and then analyze all information on

her mobile phone during the term of the loan, and to resell the data. Such borrowers may never even know what is done with their data, thanks to all-pervasive trade secrecy in the industry. Tracker apps on cell phones may allow firms to record their employees' location at all times. According to one aggrieved worker, her boss "bragged that he knew how fast she was driving at specific moments ever since she had installed [an] app on her phone."[11]

Even when such comprehensive surveillance is not consented to, AI can operate as a "prediction machine," analyzing data to make damaging inferences about individuals.[12] These inferences may be entirely unexpected, based on correlations that can only be found in vast troves of data.[13] For example, a person's proclivity to be depressed may be related to the apps or websites they visit (or how they use those apps and websites). The websites themselves may keep such data, or it may be collected by third parties with commercial or other relationships with the sites or apps.

Correlations based on how a person uses their phone or computer may be entirely unexpected. For example, a high-ranking Catholic cleric in the U.S. was recently reported to be a user of the gay dating site Grindr by journalists at the digital publication named The Pillar, based on computing data. As journalist Molly Olmsted explained:

> According to one privacy engineer who has worked on issues related to location data, Pillar (or the group that had offered CNA [the Catholic News Agency] the data back in 2018) probably purchased a data set from a data broker, which in turn had likely purchased the data from a third-party ad network that Grindr uses. Grindr itself would not have been the source of the data, but the ad network would have been given full access to the users' information as long as they agreed to Grindr's terms of services. (In 2018, Grindr, which uses highly granular location information, was found to have shared users' anonymized locations, race, sexual preferences, and even HIV status with third-party analytic firms.)[14]

Whatever your views about the cleric in this case, the generalized exposure of a person's dating practices, or other intimate inferences based on location-based data, is something a mature privacy law should prevent.

Researchers have also analyzed certain activities of people who extensively searched for information about Parkinson's disease on Bing, including their mouse movements six months before they entered those search terms.[15] Most users of the internet are probably unaware that not just what they click on, but how fast and smoothly they move their mouse to do so, can be recorded and traced by the sites they are using. The group of Bing users who searched for Parkinson's—which it is probably safe to assume is far more likely to have Parkinson's than the population as a whole—tended to have certain tremors in their mouse movements distinct from other searchers. These tremor patterns were undetectable by humans—only machine learning could distinguish the group identified to have a higher propensity to have Parkinson's, based in part on microsecond-by-microsecond differences in speed and motion of hand movement.

There is at present no widely available defense to such detection—no close-at-hand, privacy-enhancing strategy that can prevent such technologies of inference, once developed, from classifying someone as likely to develop grave illness. Perhaps a clever technologist could develop tools to "fuzz" tremor movements, or to smooth or normalize their transmission so that they do not signal abnormality. But we should not expect internet users to defend themselves against such invasive categorization by joining an arms race of deflection and rediscovery, obfuscation and clarification, encryption and decryption.[16] It is a waste of our time. And those of higher socioeconomic status have many more resources at hand to engage in such arms races, thus adding the insult of rising inequality to the injury of privacy harm.[17]

Moreover, a patchwork of weak and often underenforced privacy laws is no match for the threats posed by large-scale data processing, which can be used to either overtly or secretly rank, rate, and evaluate people, often to their detriment and unfairly.[18] Without a society-wide commitment to fair data practices, a troubling era of digital discrimination will be entrenched. The more data about signals of future distress, illness, or disability are available, the better AI will predict these conditions, enabling unscrupulous actors to take advantage of them.

II. THE IMPRACTICALITY OF CONSENT-BASED MODELS IN AN AGE OF BIG DATA

MANY SYSTEMS OF PRIVACY regulation begin with consent, shifting responsibility to data subjects to decide to whom they should grant their data, and to whom they should deny it. On the consent-based view, it is up to the data subject to vet the reliability of entities seeking access to her data and monitor their ongoing abidance with the promises they made when they obtained it. A consent-based model makes sense as part of a contractarian and propertarian view of legal order: Data are like objects, property of a person who may freely contract to give or share the data with others.[19] On this view, just as a corporation needs consent to, say, take a person's property, it needs consent to take personal data. Once a contract is struck, it governs the future exchange and use of the data.

This consent-based approach has multiple infirmities.[20] Much data arises out of observation unrestricted by even theoretical contracts. To give an example: A Google user may "consent" to data collection while using the service, but no one is asked before the cameras on Google's street-photographing cars roll down the road to grab data for its mapping service. The firm does blot out faces when publishing the images to the internet, but the principle remains: All manner of information can be garnered by microphones, cameras, and sensors trained on the public at large, and connected to particular persons, automobiles, or households via massive facial recognition databases. Aerial surveillance can trace a person's every outdoor step and movement from home and back each day, as Baltimore's recent "spy plane" litigation has revealed.[21]

Even for data that can be practically bound by contract, terms-of-service agreements are continually subject to change. These changes almost always favor service providers—despite the serious lock-in costs and reliance interests of users. If one is dealing with a monopoly or must-have service, there is no real choice but to accept their terms.[22] Even when choice is superficially available, terms of service are by and large similar and imposed as a fait accompli: The user either accepts them or does not get to use the service. Sometimes, under exceedingly gentle pressure from regulators, firms will magnanimously offer assurances about certain limits on the uses of data.[23]

However, data subjects are now under surveillance by so many firms that it is impossible for them to audit such assurances comprehensively. How can a person with a job and family to take care of try to figure out which of thousands of data controllers has information about them, has correct information, and has used it in a fair and rigorous manner? In the U.S., even the diligent will all too often run into the brick walls of trade secrecy, proprietary business methods, and malign neglect if they do so much as ask about how their data has been used, with whom it has been shared, and how it has been analyzed.[24] Europeans may make "subject access requests," but there are far too many data gathering and data processing firms for the average person to conduct reviews of their results in a comprehensive way.

The analogy between data and property also breaks down in an era of digitization, when data can be so easily copied, transferred, stored, and backed up.[25] Securing a house is a relatively easy matter compared with securing one's data. Even the most conscientious data subjects need only slip once, failing to read a key clause in terms of service, or transacting with an unreliable or insecure counterparty. Then critical, discrediting, disadvantaging, or embarrassing data about them could end up copied and recopied, populating countless databases. And even when the data subject has clearly been wronged by a data controller or processor, the judiciary may make litigation to obtain redress difficult or impossible.[26]

As data and its analysis, sharing, and inferences proliferate, the consent model becomes less and less realistic. There is simply too much for any individual to keep track of. Nor have data unions risen to the challenge, given the immense difficulty of attaining any kind of bargaining leverage vis-à-vis first-party data collectors. There are myriad data gatherers, hundreds of scoring entities,[27] and blacklists covering housing, voting, travel, and employment.[28] Even if consent-based regimes are well-administered, they can result in data gaps that impede, for instance, both medical research and opportunities for clinical care. Forced into an environment where few adverse uses of compromising data and inferences are forbidden (and where the penalties for such uses are not sufficient to deter wrongdoers), data subjects easily forego valuable opportunities (such as future AI-enabled diagnoses) or eschew low-risk chances to contribute to the public good by sharing data (for research studies).

Moreover, we cannot reasonably expect good administration of consent-based regimes in many areas. In the U.S., regulatory agencies are ill-suited to enforce others' contracts. Even when they do bring cases on the basis of deception claims, the penalties for failure to comply have frequently been dismissed as a mere cost of doing business. First Amendment defenses may also complicate any lawsuit predicated on an effort to stop the transfer or analysis of data and inferences. In Europe, while data protection authorities are empowered by law to advance the interests of data subjects via the General Data Protection Regulation (GDPR), they have in practice proven reluctant to impose the types of penalties necessary to ensure adherence to the law.[29]

III. BEYOND CONSENT: THE EX ANTE REGULATORY IMPERATIVE

One way to promote large-scale data processing that is more responsive to the public interest is to ensure that proper scrutiny occurs before the collection, analysis, and use of data. If enacted via a licensure regime, this scrutiny would enable a true industrial policy for big data, deterring misuses and thereby helping to channel AI development in more socially useful directions.[30] As AI becomes more invasive and contested, there will be increasing calls for licensure regimes. To be legislatively viable, proposals for licensure need theoretical rigor and practical specificity. What are the broad normative concerns motivating licensure? And what types of uses should be permitted?

Cognizant of these queries, some legislators and regulators have begun to develop an explicitly governance-driven approach to data.[31] While not embracing licensure, Sen. Sherrod Brown of Ohio has demonstrated how substantive limits may be enforced via licensure restrictions for large-scale data collection, analysis, and use.[32] His Data Accountability and Transparency Act would amount to a Copernican shift in U.S. governance of data, putting civil rights protection at the core of this approach to data regulation.[33] This reflects a deep concern about the dangers of discrimination against minoritized or disadvantaged groups, as well as against the "invisible

minorities" I have previously described in *The Black Box Society*.[34] For example, the mouse microtremor example mentioned above may be prevented by the Data Accountability and Transparency Act, which would forbid the calculation of the inference itself by entities that intend to discriminate based on it (or, more broadly, entities that have not demonstrated a personal or public health rationale for creating, disseminating, or using it).[35] On the other hand, the inference may be permissible as a way of conducting "public or peer-reviewed scientific, historical, or statistical research in the public interest, but only to the extent such research is not possible using anonymized data."[36] Thus, the generalizable finding may be made public, but its harmful use against an individual would be precluded by preventing a firm with no reasonable method of improving the person's health from making the inference. This avoids the "runaway data" problem I described in *Black Box Society*, where data collection and analysis initially deemed promising and helpful becomes a bane for individuals stigmatized by them.

Such assurances should enable more societal trust in vital data collection initiatives, like for health research, pandemic response, and data-driven social reform. For a chilling example of a loss of trust in a situation without such protections, we need only turn to the misuse of prescription databases in the U.S. In the 2000s, patients in the U.S. were assured that large databases of prescription drug use would be enormously helpful to them if they ended up in an emergency room away from home, since emergency doctors could have immediate access to this part of their medical record, and avoid potentially dangerous drug interactions. However, that use of the database was not immediately profitable and did not become widespread. Rather, the database became a favored information source of private insurers seeking to deny coverage to individuals on the basis of "preexisting conditions."[37] To avoid such future misuses and abuses of trust, we must develop ways of preventing discriminatory uses of personal data, and of shaping the data landscape generally, rather than continuing with a regime of post hoc, partial, and belated regulation.[38]

Sensitive to such misuses of data, ethicists have called for restrictions on certain types of AI, with a presumption that it be banned unless licensed. For example, it may be reasonable for states to develop highly specialized databases of the faces of terrorists. But to deploy such powerful technology

to ticket speeders or ferret out benefits fraud is inappropriate, like using a sledgehammer to kill a fly.[39] A rational government would not license the technology for such purposes, even if it would be entirely reasonable to do so for other purposes (for example, to prevent pandemics via early detection of infection clusters). Nor would it enable many of the forms of discrimination and mischaracterization now enabled by light-to-nonexistent regulation of large-scale data collection, analysis, and use.

The first order of business for a reformed data economy is to ensure that inaccurate, irresponsible, and damaging data collection, analysis, and use are limited. Rather than assuming that data collection, processing, and use are in general permitted, and that regulators must struggle to catch up and outlaw particular bad acts, a licensure regime flips the presumption. Under it, large-scale data collectors, brokers, and analysts would need to apply for permission for their data collection, analysis, use, and transfer (at the very least for new data practices, if older ones are "grandfathered" and thus assumed to be licensed). To that end, a stricter version of the Data Accountability and Transparency Act might eventually insist that data brokers obtain a license from the government in order to engage in the collection, sale, analysis, and use of data about identifiable people.

IV. FREE EXPRESSION CONCERNS RAISED BY THE LICENSURE OF LARGE-SCALE PERSONAL DATA COLLECTION, ANALYSIS, AND USE

Whether applied to data or the AI it powers, licensure regimes will face challenges based on free expression rights.[40] The ironies here are manifold. The classic scientific process is open, inviting a community of inquirers to build on one another's works; meanwhile, the leading corporate data hoarders most likely to be covered by the licensing regime proposed here are masters of trade secrecy, aggressively blocking transparency measures. Moreover, it is now clear that the corporate assertion of such alleged constitutional rights results in databases that chill speech and online participation.[41] It is one thing to go to a protest when security personnel watch from afar. It is quite another when the police

can immediately access your name, address, and job from a quick face scan purchased from an unaccountable private firm.

This may be one reason why the American Civil Liberties Union decisively supported the regulation of Clearview AI (a firm providing facial recognition services) under the Illinois Biometric Information Privacy Act (BIPA), despite Clearview's insistence (to courts and the public at large) that it has a First Amendment right to gather and analyze data unimpeded by BIPA. If unregulated, the firm's activities seem far more likely to undermine a robust public sphere than to promote it. Moreover, even if its data processing were granted free expression protections, such protections may be limited by "time, place, and manner" restrictions. In that way, the licensure regime I am proposing is much like permit requirements for parades, which recognize the need to balance the parade organizers' and marchers' free expression rights against the public need for safe and orderly streets. Given the privacy and security concerns raised by mass data collection, analysis, and use, restrictions on data practices thus may be subject to only intermediate scrutiny in the U.S.[42] Even more sensible is the Canadian rejection of the data aggregator's free expression claim tout court.[43]

When an out-of-control data gathering industry's handiwork can be appropriated by both government and business decision-makers, data and inferences reflect both knowledge and power: They are descriptions of the world that also result in actions done within it. They blur the boundary between speech and conduct, observation and action, in ways that law can no longer ignore. Mass data processing is unlike the ordinary language (or "natural language") traditionally protected by free expression protections. Natural language is a verbal system of communication and meaning-making. I can state something to a conversation partner and hope that my stated (and perhaps some unstated) meanings are conveyed to that person. By contrast, in computational systems, data are part of a project of "operational language"; their entry into the system produces immediate effects. As Mark Andrejevic explains in *Automated Media*, there is no interpretive gap in computer processing of information.[44] The algorithm fundamentally depends on the binary (1 or 0), supplemented by the operators "and, or, not." In Andrejevic's words, "machine 'language' … differs from human language precisely because it is non-representational. For the machine, there

is no space between sign and referent: there is no 'lack' in a language that is complete unto itself. In this respect, machine language is 'psychotic' ... [envisioning] the perfection of social life through its obliteration."[45] This method of operation is so profoundly different than human language—or the other forms of communication covered by free expression protections—that courts should be exceptionally careful before extending powerful "rights to speak" to the corporate operators of computational systems that routinely abridge human rights to privacy, data protection, and fair and accurate classification.

Unregulated AI is always at risk of distorting reality. Philosophers of social science have explained the limits and constraints algorithmic processing has imposed on social science models and research.[46] Scholars in critical data studies have exposed the troubling binaries that have failed to adequately, fairly, and humanely represent individuals. For example, Os Keyes has called data science a "profound threat for queer people" because of its imposition of gender binaries on those who wish to escape them (and who seek societal acceptance of their own affirmation of their gender).[47] In this light, it may well be the case that an entity should only process data about data subjects' gender (and much else) if it has been licensed to do so, with licensure authorities fully cognizant of the concerns that Keyes and other critical data scholars have raised.

The shift to thinking of large-scale data processing as a privilege, instead of as a right, may seem jarring to American ears, given the expansion of First Amendment coverage over the past century.[48] However, even in the U.S. it is roundly conceded that there are certain particularly sensitive pieces of "information" that cannot simply be collected and disseminated. A die-hard cyberlibertarian or anarchist may want to copy and paste bank account numbers or government identification numbers onto anonymous websites, but that is illegal because complex sociotechnical systems like banks and the Social Security Administration can only function on a predicate of privacy and informational control.[49] We need to begin to do the same with respect to facial recognition and other biometrics, and to expand this caution with respect to other data that may be just as invasive and stigmatizing. Just as there is regulation of federally funded human subjects research, similar patterns of review and limitation must apply to the new forms of human classification and manipulation now enabled by massive data collection.[50]

A licensure regime for big data analytics also puts some controls on the speed and ubiquity of the correlations such systems can make. Just as we may want to prevent automated bots from dominating forums like Twitter, we can and should develop a societal consensus toward limiting the degree to which automated correlations of often biased, partial, and secret data influence our reputations and opportunities.[51]

This commitment is already a robust part of finance regulation. For example, when credit scores are calculated, the Fair Credit Reporting Act imposes restrictions on the data that can affect them.[52] Far from being a forbidden content-based restriction on the "speech" of scoring, such restrictions are vital to a fair credit system.[53] The Equal Credit Opportunity Act takes the restrictions further regarding a creditor's scoring system.[54] Such scoring systems may not use certain characteristics—such as race, sex, gender, marital status, national origin, religion, or receipt of public assistance—as a factor regarding a customer's creditworthiness.[55] Far from being a relic of the activist 1970s, restrictions like this are part of contemporary efforts to ensure a fairer credit system. [56]

European examples abound as well. In Germany, the United Kingdom, and France, agencies cannot use ethnic origin, political opinion, trade union membership, or religious beliefs when calculating credit scores.[57] Germany and the United Kingdom also prohibit the use of health data, while France allows the use of health data in credit score calculations.[58] Such restrictions might be implemented as part of a licensure regime for use of AI-driven propensity scoring in many fields. For example, authorities may license systems that credibly demonstrate to authorized testing and certification bodies that they do not process data on forbidden grounds, while denying a license to those that do.

Moreover, credit scores themselves feature as forbidden data in some other determinations. For example, many U.S. states prevent them from being used by employers.[59] California, Hawaii, and Massachusetts ban the use of credit scoring for automobile insurance.[60] A broad coalition of civil rights and workers' rights groups reject these algorithmic assessments of personal worth and trustworthiness.[61] The logical next step for such activism is to develop systems of evaluation that better respect human dignity and social values in the construction of actionable reputations—those with

direct and immediate impact on how we are classified, treated, and evaluated. For example, many have called for the nationalization of at least some credit scores.[62] Compared with that proposal, a licensure regime for such algorithmic assessments of propensity to repay is moderate.

To be sure, there will be some difficult judgment calls to be made, as in the case with any licensure regime. But size-based triggers can blunt the impact of licensure regimes on expression, focusing restrictions on firms with the most potential to cause harm. These firms are so powerful that they are almost governmental in their own right.[63] The EU's Digital Services Act proposal, for example, includes obligations that would only apply to platforms that reach 10 percent of the EU population (about 45 million people).[64] The Digital Markets Act proposal includes obligations that would only apply to firms that provide "a core platform service that has more than 45 million monthly active end users established or located in the Union and more than 10,000 yearly active business users established in the Union in the last financial year."[65] In the U.S., the California Consumer Privacy Act applies to companies that have data on 50,000 California residents.[66] Many U.S. laws requiring security breach notifications generally trigger at around 500-1,000 records breached.[67] In short, a nuanced licensing regime can be developed that is primarily aimed at the riskiest collections of data, and only imposes such obligations (or less rigorous ones) on smaller entities as the value and administrability of requirements for larger firms is demonstrated.

V. CONCLUSION

One part of the "grand bargain for big data" I outlined in 2013, followed by the "redescription of health privacy" I proposed in 2014, is a reorientation of privacy and data protection advocacy.[68] The state, its agencies, and the corporations they charter only deserve access to more data about persons if they can demonstrate that they are actually using that data to advance human welfare. Without proper assurances that the abuse of data has been foreclosed, citizens should not accede to the large-scale data grabs now underway.

Not only ex post enforcement but also ex ante licensure is necessary to ensure that data are only collected, analyzed, and used for permissible purposes. This article has sketched the first steps toward translating the general normative construct of a "social license" for data use into a specific licensure framework. Of course, more conceptual work remains to be done, both substantively (elaborating grounds for denying a license) and practically (to estimate the resources needed to develop the first iteration of the licensing proposal).[69] The consent model has enjoyed the benefits of such conceptual work for decades; now it is time to devote similar intellectual energy to a licensing model.

Ex ante licensure of large-scale data collection, analysis, use, and sharing should become common in jurisdictions committed to enabling democratic governance of personal data. Defining permissible purposes for the licensure of large-scale personal data collection, analysis, use, and sharing will take up an increasing amount of time for regulators, and law enforcers will need new tools to ensure that regulations are actually being followed. The articulation and enforcement of these specifications will prove an essential foundation of an emancipatory industrial policy for AI.

NOTES

1 In this article, I collectively refer to the collection, analysis, transfer, and use of data as "data practices." The analysis of one set of data can create a new set of data, or inferences; for my purposes, all this follow-on development of data and inferences via analysis is included in the term analysis itself.

2 For earlier examples of this kind of move to supplement ex post regulation with ex ante licensure, see Saule T. Omarova, *License to Deal: Mandatory Approval of Complex Financial Products*, 90 Wash. Univ. L. Rev. 63, 63 (2012); Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83 (2017); Frank Pasquale, The Black Box Society 181 (2015). The Federal Communications Commission's power to license spectrum and devices is also a useful precedent here—and one reason my epigraph for this piece gestures to the work and views of one of the most influential FCC commissioners in U.S. history, Newton Minow. Like the airwaves, big data may usefully be considered as a public resource. Salome Viljoen, *Democratic Data: A Relational Theory For Data Governance*, Yale L. J. (forthcoming, 2021).

3 Siddharth Venkataramakrishnan, *Top researchers condemn 'racially biased' face-based crime prediction*, Fin. Times (June 24, 2020), https://www.ft.com/content/aaa9e654-c962-46c7-8dd0-c2b4af932220 [https://perma.cc/AZU5-VLPD] ("More than 2,000 leading academics and researchers from institutions including Google, MIT, Microsoft and Yale have called on academic journals to halt the publication of studies claiming to have used algorithms to predict criminality. The nascent field of AI-powered 'criminal recognition' trains algorithms to recognise complex patterns in the facial features of people categorised by whether or not they have previously committed crimes."). For more on the problems of face-focused prediction of criminality by AI, see Frank Pasquale, *When Machine Learning is Facially Invalid*, 61 Comm'ns ACM 25, 25 (Sept. 2018).

4 *See also* Daten ethik kommission [Data Ethics Commission of the Federal Government of Germany], *Opinion of the Data Ethics Commission* (2019), 195 (calling for "Preventive official licensing procedures for high-risk algorithmic systems"). The DEC observes that, "[I]n the case of algorithmic systems with regular or appreciable (Level 3) or even significant potential for harm (Level 4), in addition to existing regulations, it would make sense to establish licensing procedures or preliminary checks carried out by supervisory institutions in order to prevent harm to data subjects, certain sections of the population or society as a whole." *Id*. Such licensing could also be promulgated by national authorities to enforce the European Union's proposed AI Act. Frank Pasquale & Gianclaudio Malgieri, *Here's a Model for Reining in AI's Excesses*, N.Y. Times, Aug. 2, 2021, at A19.

5 Gina Kolata, *The First Word on Predicting Alzheimer's*, N.Y. Times, Feb. 2, 2021, at D3.

6 Pasquale, *supra* note 2, at 149; Frank Pasquale, *Promoting Data for Well-Being While Minimizing Stigma: Foundations of Equitable AI Policy for Health Predictions*, *in* Digital Dominance (Martin Moore & Damian Tambini, eds., Oxford University Press, 2021).

7 For more on this analogy between big data-driven health prediction and human subjects research, see James Grimmelmann, *The Law and Ethics of Experiments on Social Media Users*, 13 Colo. Tech. L. J. 219 (2015); Frank Pasquale, P*rivacy, Autonomy, and Internet Platforms*, *in* Privacy In The Modern Age: The Search For Solutions (Marc Rotenberg et al. eds., 2015).

8 Theodore Rostow, *What Happens When an Acquaintance Buys Your Data?: A New Privacy Harm in the Age of Data Brokers*, 34 Yale J. on Reg. 667 (2017).

9 Brent Mittelstadt, *From Individual to Group Privacy in Biomedical Big Data*, *in* Big Data, Health Law, and Bioethics 175 (I. Glenn Cohen et al. eds., 2018); Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2 Colum. Bus. L. Rev. (2019).

10 Aaron Chou, *What's In The "Black Box"? Balancing Financial Inclusion and Privacy in Digital Consumer Lending*, 69 Duke L. J. 1183, 1192 (2020).

11 James Vincent, *Woman fired after disabling work app that tracked her movements 24/7*, Verge (May 13, 2015, 7:01 AM), https://www.theverge.com/2015/5/13/8597081/worker-gps-fired-myrna-

arias-xora.

12 Ajay Agrawal et al., Prediction Machines: The Simple Economics of Artificial Intelligence (2018).

13 Eric Horvitz & Deirdre Mulligan, *Data, privacy, and the greater good*, Science, July 17, 2015, at 253.

14 Molly Olmsted, *A Prominent Priest Was Outed for Using Grindr. Experts Say It's a Warning Sign*, Slate (July 21, 2021, 7:03 PM), https://slate.com/technology/2021/07/catholic-priest-grindr-data-privacy.html [https://perma.cc/DN5H-J8NA].

15 Ryen W. White et al., *Detecting Neurogenerative Disorders from Web Search Signals*, 1 NPJ Digit. Med. 1 (Apr. 23, 2018), https://www.nature.com/articles/s41746-018-0016-6.pdf [https://perma.cc/9KDK-VBGZ]. In this case, the source of the information was clear: Microsoft itself, which operates Bing, permitted the researchers to study anonymized databases. For an analysis of the import of such data in the U.S., where it is now well beyond the scope of the privacy and security protections guaranteed pursuant to the Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health (HITECH) Act), see National Committee on Vital and Health Statistics, Subcommittee on Privacy, Confidentiality, and Security, *Health Information Privacy Beyond HIPAA* (2019), at https://ncvhs.hhs.gov/wp-content/uploads/2019/07/Report-Framework-for-Health-Information-Privacy.pdf [https://perma.cc/A5JW-QLW8].

16 As I elaborate in a recent book, one critical goal of technology policy should be stopping such arms races. Frank Pasquale, New Laws of Robotics (2020); *see also* Frank Pasquale, *Paradoxes of Privacy in an Era of Asymmetrical Social Control*, *in* Big Data, Crime and Social Control (Aleš Završnick ed., 2018) (on the encryption arms race). As Jathan Sadowski has argued, "cyberhygiene" arguments all too easily degenerate into victim-blaming. Jathan Sadowski, Too Smart (2019).

17 Of course, in the right hands, the data could be quite useful. We may want our doctors to access such information, but we need not let banks, employers, or others use it. That is one foundation of the licensing regime I will describe in Part III: ensuring persons can generally presume data associated with them is being used to advance their well-being, rather than to stigmatize or exclude them.

18 For a powerful critique of extant privacy laws in the U.S. and Europe, see Julie Cohen, Between Truth and Power (2019).

19 On the fallacies inherent in this model in the context of health privacy, see Barbara J. Evans, *Much Ado About Data Ownership*, 25 Harv. J. L. & Tech. 70 (2011).

20 Julie Cohen, *Turning Privacy Inside Out*, 20 Theoretical Inquiries L. 1, 1 (2019) ("[P]rivacy's most enduring institutional failure modes flow from its insistence on placing the individual and individualized control at the center."); Bart Willem Schermer et al., *The Crisis of Consent: How Stronger Legal Protection May Lead to Weaker Consent in Data Protection*, 16 Ethics and Info. Tech. 171 (2014); Gabriela Zanfir-Fortuna, *Forgetting About Consent: Why the Focus Should Be on 'Suitable Safeguards' in Data Protection Law* (May 10, 2013) (unpublished working paper) (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2261973 [https://perma.cc/DU9X-97ND]).

21 Leaders of a Beautiful Struggle v. Balt. Police Dep't, 1:20-cv-00929-RDB (4th Cir., June 24, 2021). A sharply divided Fourth Circuit Court of Appeals ruled the program unconstitutional.

22 Frank Pasquale, *Privacy, Antitrust, and Power*, 20 Geo. Mason L. Rev. 1009 (2013); Andreas Mundt, *Bundeskartellamt prohibits Facebook from combining user data from different sources*, *Bundeskartellamt* (Feb. 7, 2019), https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html [https://perma.cc/X6US-RL7A] ("In view of Facebook's superior market power, an obligatory tick on the box to agree to the company's terms of use is not an adequate basis for such intensive data processing. The only choice the user has is either to accept the comprehensive combination of data or to refrain from using the social network. In such a difficult situation the user's choice cannot be referred to as voluntary consent.").

23 Ari Ezra Waldman, Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power (2021).

24 Even in the health care system, where access to such information is supposed to be guaranteed by federal health privacy laws, patients find considerable barriers to the exercise of their rights.

25 For a sophisticated response to this problem, see Václav Janeček & Gianclaudio Malgieri, *Commerce in Data and the Dynamically Limited Alienability Rule*, 21 Ger. L. J. 924 (2020).

26 Daniel J. Solove & Danielle Keats Citron, *Standing and Privacy Harms: A Critique of* TransUnion v. Ramirez, 101 B. U. L. Rev. Online 62 (2021).

27 Pam Dixon & Bob Gellman, *The Scoring of America* (World Policy Forum, Apr. 2, 2014), http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf [https://perma.cc/GP6J-L75J]; Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1, 31 (2014).

28 Margaret Hu, *Big Data Blacklisting*, 67 Fla. L. Rev. 1735 (2016).

29 Adam Satariano, *Europe's Privacy Law Hasn't Shown Its Teeth*, N.Y. Times, Apr. 28, 2020, at B1.

30 In this way, my proposals here are an extension of the ideas I develop in Frank Pasquale, *Data Informed Duties for AI Development*, 119 Colum. L. Rev. 1917, 1917 (2019) ("Law should help direct—and not merely constrain—the development of artificial intelligence (AI). One path to influence is the development of standards of care both supplemented and informed by rigorous regulatory guidance.").

31 For example, the European Data Protection Board is exploring certification. Eur. Data Prot. Bd., *Guidelines 1/2018 on certification and identifying certification criteria in accordance with Articles 42 and 43 of the Regulation*, Version 3.0 (June 4, 2019), https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-12018-certification-and-identifying_en [https://perma.cc/Q365-M99F] ("Before the adoption of the GDPR, the Article 29 Working Party established that certification could play an important role in the accountability framework for data protection. In order for certification to provide reliable evidence of data protection compliance, clear rules setting forth requirements for the provision of certification should be in place. Article 42 of the GDPR provides the legal basis for the development of such rules.").

32 In future work, I hope to compare Brown's proposal with the GDPR's definition of "legitimate purposes." Under the GDPR, "Personal data shall be … collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes." Chris Jay Hoofnagle et al., *The European Union general data protection regulation: what it is and what it means*, 28 Info. & Comms. Tech. L. 65, 77 n. 82 (2019) (quoting Council Directive 2016/679 O.J. (L 119) art. 5(1)(b) (General Data Protection Regulation)).

33 Press Release, Sherrod Brown U.S. Sen. for Ohio, *Brown Releases New Proposal That Would Protect Consumers' Privacy from Bad Actors* (June 18, 2020), (https://www.brown.senate.gov/newsroom/press/release/brown-proposal-protect-consumers-privacy [https://perma.cc/74HR-KJLA]).

34 Pasquale, *supra* note 2. For a deep development of the invisible minorities idea as a right to avoid being stigmatized by certain unreasonable inferences about groups, see Sandra Wachter & Brent Mittelstadt, *supra* note 9. *See also* Gianclaudio Malgieri & Jędrzej Niklas, *Vulnerable Data Subjects*, Comput. L. & Sec. Rev. 37 (July 2020).

35 Data Accountability and Transparency Act (DATA Act), S. 20719, 116th Cong. § 102(b)(4) (as proposed to the Senate, 2020) [hereinafter DATA Act]. The proposed act states that data aggregators "shall not collect, use, or share, or cause to be collected, used, or shared, any personal data unless the aggregator can demonstrate that such personal data is strictly necessary to carry out a permissible purpose under section 102." *Id.* at § 101. It also states that "a data aggregator shall not … derive or infer data from any element or set of personal data."

36 *Id.* at § 102(a)(3). For European efforts to define a similar category, see Eur. Data Prot. Supervisor, *Preliminary Opinion on data protection and scientific research*, Eur. Data. Prot. Supervisor (Jan. 6, 2020), https://edps.europa.eu/data-protection/our-work/publications/opinions/preliminary-opinion-data-protection-and-scientific_en [https://perma.cc/4T3H-7W9W].

37 Chad Terhune, *They Know What's in Your Med-*

icine Cabinet, Bloomberg Businessweek (July 23, 2008, 12:00 AM), https://www.bloomberg.com/news/articles/2008-07-22/they-know-whats-in-your-medicine-cabinet [https://perma.cc/3ZY4-ZVMX]. Of course, the guaranteed issue provisions and ban on preexisting condition limitations in the 2010 Affordable Care Act (ACA) made such practices much less menacing to most consumers. However, the ACA could easily be repealed, or declared null and void by an activist Supreme Court. The rise of authoritarianism in the U.S. should further caution us to understand that no such rights (except of course those of the party in power and its allies) are permanently entrenched.

38 The proposed DATA Act's "Prohibition On Discriminatory Use of Personal Data" is a method for shaping data collection, analysis, and use in a democratically accountable and forward-thinking way. DATA Act § 104. ("It is unlawful for a data aggregator to collect, use, or share personal data for ... commercially contracting for housing, employment, credit, or insurance in a manner that discriminates against or otherwise makes the opportunity unavailable or ordered on different terms on the basis of a protected class."). As defined by the DATA Act, "protected class" includes classifications based on "biometric information," which would cover hand-motion monitoring (and many other, more remote forms of data collection and classificatory inference). "Protected class" is defined as "actual or perceived race, color, ethnicity, national origin, religion, sex, gender, gender identity, sexual orientation, familial status, biometric information, lawful source of income, or disability of an individual or group of individuals." DATA Act § 3(20).

39 For an example of other such potential excessive uses, see Robert Pear, *On Disability and on Facebook? Uncle Sam Wants to Watch What You Post*, N.Y. Times, (Mar. 10, 2019), https://www.nytimes.com/2019/03/10/us/politics/social-security-disability-trump-facebook.html [https://perma.cc/7FZ4-MFLK].

40 These rights claims will be particularly salient in the U.S., whose courts have expanded the scope of the First Amendment to cover many types of activity that would not merit free expression elsewhere, or would merit much less intense free expression protection, given the importance of competing rights to privacy, security, and data protection. On the general issue of data's categorization as speech, see Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. Davis L. Rev. 1183 (2016); Jane Bambauer, *Is Data Speech?*, 66 Stan. L. Rev. 57 (2014); Paul M. Schwartz, *Free Speech vs. Information Privacy: Eugene Volokh's First Amendment Jurisprudence*, 52 Stan. L. Rev. 1559 (2000); James M. Hilmert, *The Supreme Court Takes on the First Amendment Privacy Conflict and Stumbles: Bartnicki v. Vopper, the Wiretapping Act, and the Notion of Unlawfully Obtained Information*, 77 Ind. L. J. 639 (2002); Eric B. Easton, *Ten Years After: Bartnicki v. Vopper as a Laboratory for First Amendment Advocacy and Analysis*, 50 U. Louisville L. Rev. 287 (2011).

41 Johanna Gunawan et al., *The COVID-19 Pandemic and the Technology Trust Gap*, 51 Seton Hall L. Rev. 1505 (2021).

42 ACLU v. Clearview AI, Case 20 CH 4353, (Ill. Cir., Aug. 27, 2021), at 10 ("BIPA's speaker-based exemptions do not appear to favor any particular viewpoint. As BIPA's restrictions are content neutral, the Court finds that intermediate scrutiny is the proper standard.").

43 *Joint investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta*, PIPEDA Findings #2021-001, para. 67, https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/ [https://perma.cc/XN8W-LKV8] ("Clearview has neither explained nor demonstrated how its activities constitute the expression of a message relating to the pursuit of truth, participation in the community or individual self-fulfillment and human flourishing.").

44 Mark Andrejevic, Automated Media (2020).

45 *Id.* at 72.

46 Paul Erickson et al., How Reason Almost Lost Its Mind: The Strange Career of Cold War Rationality (2013); S.M. Amadae, *Game Theory, Cheap Talk and Post-Truth Politics: David Lewis vs. John Searle on reasons for truth-telling*, 48 J. Theory

Soc. Behav. 306 (2018).

47 Os Keyes, *Counting the Countless*, Real Life (Apr. 8, 2019), https://reallifemag.com/counting-the-countless/ [https://perma.cc/7M9J-4XFK].

48 Note that the DATA Act has an exception for "de minimis" collection, analysis, and use: "Any person that collects, uses, or shares an amount of personal data that is not de minimis; and does not include an individual who collects, uses, or shares personal data solely for personal reasons." DATA Act, § 3(8)(A)-(B). The "large-scale" proviso of the licensure regime proposed in this work is also meant to shield smaller players, but on a larger scale.

49 For a broader argument on the limits of First Amendment protection for operational code, see David Golumbia, *Code is Not Speech* (Apr. 13, 2016) (unpublished draft) (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2764214 [https://perma.cc/G8UG-XMAQ]).

50 For an analysis of the analogy between many forms of big data processing and experiments that are clearly deemed human subjects research, see James Grimmelmann, *The Law and Ethics of Experiments on Social Media Users*, 13 Colo. Tech. L. J. 219 (2015), https://ctlj.colorado.edu/wp-content/uploads/2015/08/Grimmelman-final.pdf [https://perma.cc/7N6K-JJHG].

51 On policy rationales for limiting automated bot speech, see Frank Pasquale, *Preventing a Posthuman Law of Freedom of Expression*, *in* The Perilous Public Sphere (David Pozen ed., 2020).

52 U.S. Fair Credit Reporting Act (FCRA) § 609, 15 U.S.C. § 1681(g) (2011).

53 The FCRA provides further language limiting what information may by contained in a consumer report. 15 U.S.C. 1681(c) (2011). Consumer reports cannot contain: Title 11 cases over 10 years old; civil suits, judgments, or arrest records over seven years old; paid tax liens over seven years old; accounts placed for collection or charged to profit and loss over seven years old; or any other adverse information, other than criminal convictions, over seven years old. These restrictions have not been successfully challenged as content-based restrictions under the First Amendment.

54 A creditor is defined by the Equal Credit Opportunity Act as those who "extend, renew, or continue credit." 15 U.S.C. § 1691(a)(e) (2010).

55 15 U.S.C. § 1691(a).

56 In New York, legislation was passed that bans consumer reporting agencies and lenders from using a consumer's social network to determine creditworthiness. The bill specifically bans companies from using the credit scores of people in an individual's social network as a variable in determining their credit score. Keshia Clukey, *Social Networks Can't Go Into Credit Decisions Under N.Y. Ban*, Bloomberg L. (Nov. 25, 2019, 5:13 PM), https://news.bloomberglaw.com/banking-law/social-networks-cant-go-into-credit-decisions-under-n-y-ban[https://perma.cc/LLA7-FMXB].

57 Nicola Jentzsch, Financial Privacy, An International Comparison of Credit Reporting Systems (2007).

58 *Id*. The same restriction applies in the U.S. "A consumer reporting agency shall not furnish ... a consumer report that contains medical information (other than medical contact information treated in the manner required under section 1681(c)(a)(6) of this title) about a consumer, unless—the consumer affirmatively consents, ... if furnished for employment purposes, ... the information is relevant to the process or effect[s] the employment or credit transaction, ... the information to be furnished pertains solely to transactions, accounts, or balances relating to debts arising from the receipt of medical services, products, or devises, ... a creditor shall not obtain or use medical information ... in connection with any determination of the consumer's eligibility, or continued eligibility, for credit." Fair Credit Reporting Act, 15 U.S.C. § 1681(b)(g) (2020).

59 *State Laws Limiting Use of Credit Information for Employment*, Microbilt (2017), https://www.microbilt.com/Cms_Data/Contents/Microbilt/Media/Docs/MicroBilt-State-Laws-Limiting-Use-of-Credit-Information-For-Employment-Version-1-1-03-01-17-.pdf [https://perma.cc/9LSS-GLXZ].

60 *Id*.

61 Assembly Floor Analysis: AB-22 Employment: credit reports, http://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=201120120AB22 [https://perma.cc/XPQ9-QM8U]

(last visited May 13, 2021). Groups include unemployed people, low-income communities, communities of color, women, domestic violence survivors, families with children, divorced individuals, and those with student loans and/or medical bills. N.Y.C. Comm'n on Hum. Rts., *Stop Credit Discrimination in Employment Act: Legal Enforcement Guidance* (N.Y.C. Comm'n on Hum. Rts. 2015), https://www1.nyc.gov/site/cchr/law/stop-credit-discrimination-employment-act.page [https://perma.cc/4TSX-8ECH].

62 McKenna Moore, *Biden wants to change how credit scores work in America*, Fortune (Dec. 18, 2020, 11:27 AM), https://fortune.com/2020/12/18/biden-public-credit-agency-economic-justice-personal-finance-racism-credit-scores-equifax-transuion-experian-cfpb/ [https://perma.cc/5Y5P-9NLS]; Amy Traub, *Establish a Public Credit Registry*, Demos (Apr. 3, 2019), https://www.demos.org/policy-briefs/establish-public-credit-registry [https://perma.cc/R998-DA6G]; The Biden Plan for Investing In Our Communities Through Housing, https://joebiden.com/housing/ (last visited July 12, 2021).

63 Frank Pasquale, *From Territorial to Functional Sovereignty: The Case of Amazon*, LPE Project (Dec. 6, 2017), https://lpeproject.org/blog/from-territorial-to-functional-sovereignty-the-case-of-amazon/ [https://perma.cc/52YQ-5RBK].

64 *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act)*, at 3, COM (2020) 825 final (Dec. 15, 2020) ("The operational threshold for service providers in scope of these obligations includes those online platforms with a significant reach in the Union, currently estimated to be amounting to more than 45 million recipients of the service. This threshold is proportionate to the risks brought by the reach of the platforms in the Union; where the Union's population changes by a certain percentage, the Commission will adjust the number of recipients considered for the threshold, so that it consistently corresponds to 10% of the Union's population."); *Id*. at 31 ("Such significant reach should be considered to exist where the number of recipients exceeds an operational threshold set at 45 million, that is, a number equivalent to 10% of the Union population. The operational threshold should be kept up to date through amendments enacted by delegated acts, where necessary."). Such thresholds reflect a risk-focused model of regulation commended by the German Data Ethics Commission. Data Ethics Comm'n, Fed. Gov't Ger., *Opinion of the Data Ethics Commission* (2019), 177.

65 *Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)*, at 36–37, COM (2020) 842 final (Dec. 15, 2020) ("A provider of core platform services shall be presumed [an important gateway for business users to reach end users] where it provides a core platform service that has more than 45 million monthly active end users established or located in the Union and more than 10,000 yearly active business users established in the Union in the last financial year.").

66 Cal. Civ. Code § 1798.140(c)(1)(B) (West 2020) (covering any business that "[a]lone or in combination, annually buys, receives for the business's commercial purposes, sells, or shares for commercial purposes, alone or in combination, the personal information of 50,000 or more consumers, households, or devices").

67 *See, e.g.*, 16 C.F.R. § 318.5(b)–(c) ("A vendor of personal health records or PHR related entity shall provide notice to prominent media outlets serving a State or jurisdiction, following the discovery of a breach of security, if the unsecured PHR identifiable health information of 500 or more residents of such State or jurisdiction is, or is reasonably believed to have been, acquired during such breach."); Security Breach Notification Laws, https://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx [https://perma.cc/BS39-J2RE] (last visited May 13, 2021) (36 states set notification thresholds at 500 or 1,000).

68 Frank Pasquale, *Grand Bargains for Big Data: The Emerging Law of Health Information*, 72 Md. L. Rev. 682 (2013); Frank Pasquale, *Redescribing Health Privacy: The Importance of Information Policy*, 14 Hous. J. Health L. & Pol'y 95 (2014).

69 To provide the proper level of resources, the "self-funding agency" model is useful. Certain financial and medical regulators are funded in part via fees paid by regulated entities that must apply to

engage in certain activities. For example, fees paid pursuant to the Prescription Drug User Fee Act (PDUFA) fund the Food and Drug Administration (which essentially licenses drugs for sale in the U.S.). For background on this act and its amendments, see Prescription Drug User Fee Amendments, https://www.fda.gov/industry/fda-user-fee-programs/prescription-drug-user-fee-amendments [https://perma.cc/5THX-NTKD] (last updated Aug. 25, 2021).

About the Author

Frank Pasquale is a professor of law at Brooklyn Law School, an affiliate fellow at the Yale Information Society Project, and the Minderoo High Impact Distinguished Fellow at the AI Now Institute. He is also the chairman of the Subcommittee on Privacy, Confidentiality, and Security of the National Committee on Vital and Health Statistics at the U.S. Department of Health and Human Services. Pasquale is an expert on the law of artificial intelligence, algorithms, and machine learning, and author of *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Harvard University Press, 2020). His widely cited book, *The Black Box Society* (Harvard University Press, 2015), develops a social theory of reputation, search, and finance, and promotes pragmatic reforms to improve the information economy, including more vigorous enforcement of competition and consumer protection law. *The Black Box Society* has been reviewed in Science and Nature, published in several languages, and its fifth anniversary of publication has been marked with an international symposium in Big Data & Society.

© 2021, Frank Pasquale.

Acknowledgments

I wish to thank David Baloche, Jameel Jaffer, Margot Kaminski, Amy Kapczynski, Gianclaudio Malgieri, Rafi Martina, Paul Ohm, Paul Schwartz, and Ari Ezra Waldman for very helpful comments on this work. I, of course, take responsibility for any faults in it. I also thank the Knight First Amendment Institute and the Law and Political Economy Project for the opportunity to be in dialogue on these critical issues.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Erik Carter



KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY