NEW YORK UNIVERSITY
LAW REVIEW

VOLUME 94 OCTOBER 2019 NUMBER 4

SYMPOSIUM

SAFE SHARING SITES

LISA M. AUSTIN† & DAVID LIE‡

In this Article we argue that data sharing is an activity that sits at the crossroads of privacy concerns and the broader challenges of data governance surrounding access and use. Using the Sidewalk Toronto "smart city" proposal as a starting point for discussion, we outline these concerns to include resistance to data monopolies, public control over data collected through the use of public infrastructure, public benefit from the generation of intellectual property, the desire to broadly share data for innovation in the public interest, social—rather than individual—surveillance and harms, and that data use be held to standards of fairness, justice, and accountability. Data sharing is sometimes the practice that generates these concerns and sometimes the practice that is involved in the solution to these concerns.

Our safe sharing site approach to data sharing focuses on resolving key risks associated with data sharing, including protecting the privacy and security of data subjects, but aims to do so in a manner that is independent of the various legal contexts of regulation and governance. Instead, we propose that safe sharing sites connect with these different contexts through a legal interface consisting of a registry that provides transparency in relation to key information that supports different forms of regulation. Safe sharing sites could also offer assurances and auditability regarding the data sharing, further supporting a range of regulatory interventions. It is therefore not an alternative to these interventions but an important tool that can enable effective regulation.

A central feature of a safe sharing site is that it offers an alternative to the strategy of de-identifying data and then releasing it, whether within an "open data" context or

† Professor, Chair in Law and Technology, University of Toronto, Faculty of Law.

‡ Professor, Department of Electrical and Computer Engineering, University of Toronto. We would like to thank Margaret Kwoka, Yafit Lev-Aretz, the participants of the symposium *Data Law in a Global Digital Economy*, 94 N.Y.U. L. REV. 581 (2019), the participants of the Guarini Colloquium, and the staff at the Information and Privacy Commissioner of Ontario for their many helpful comments on earlier versions of this project. We would also like to thank Mathew Elder for his excellent research assistance. All errors remain ours. Copyright © 2019 by Lisa M. Austin & David Lie.

in a more controlled environment. In a safe sharing site, computations may be performed on the data in a secure and privacy-protective manner without releasing the raw data, and all data sharing is transparent and auditable. Transparency does not mean that all data sharing becomes a matter of "public" view, but rather that there is the ability to make these activities visible to organizations and regulators in appropriate circumstances while recognizing the potential confidentiality interests in data uses.

In this way, safe sharing sites facilitate data sharing in a manner that manages the complexities of sharing while reducing the risks and enabling a variety of forms of governance and regulation. As such, the safe sharing site offers a flexible and modular piece of legal-technical infrastructure for the new economy.

Introduction

Described as a "21st-century battle over privacy," the proposed Sidewalk Toronto "smart city" is the controversial site of a data-driven reimagining of civic space.[1] Although the privacy issues associated with urban data collection have garnered much attention, the question of who will control access to the data and determine data uses is equally important. The controversy generated by Sidewalk Toronto highlights some of the broader emerging data governance issues surrounding access and use, including resistance to data monopolies; public control over data collected through the use of public

[1] *Sidewalk Lab's Vision and Your Data Privacy: A Guide to the Saga on Toronto's Waterfront*, Globe & Mail (Dec. 7, 2018), https://www.theglobeandmail.com/canada/toronto/article-sidewalk-labs-quayside-toronto-waterfront-explainer; Sidewalk Toronto, https://sidewalktoronto.ca (last visited Jan. 16, 2019) (describing Sidewalk Toronto as a "joint effort between Waterfront Toronto and Alphabet's Sidewalk Labs" that will develop a section of Toronto's waterfront into a "smart city").

infrastructure; public benefit from the generation of intellectual property; the desire to broadly share data for innovation in the public interest; social—rather than individual—surveillance and harms; and adherence of data to standards of fairness, justice, and accountability.[2]

The Sidewalk Toronto proposal includes two core elements to address some of these developing challenges. The first element, addressing the concerns regarding data monopolies, is that nobody will own the data. Instead, the data will be "open" by default and the de-identified data will be shared as broadly as possible. The second element, addressing the broad range of data use concerns, is to create a "Civic Data Trust" that can manage urban data in the public interest.[3] In this Article, we question the first element and propose an alternative—what we call a "safe sharing site"—that can work with different models of data governance, including data trusts.

Making de-identified data freely and publicly available raises privacy concerns because of the re-identification risks. However, current methods used to mitigate re-identification risks reduce accuracy of the data. Less accurate data can undermine efforts to further innovation and competition, since some of the data's uses require accuracy that cannot be achieved under current de-identification methods.[4]

[2] *See* Kate Allen, *AI Pioneer Urges Toronto to Back Ethical Use of Artificial Intelligence*, Star (Apr. 11, 2019), https://www.thestar.com/news/gta/2019/04/11/ai-pioneer-urges-toronto-to-back-ethical-use-of-artificial-intelligence.html; Andrew Clement, *Sidewalk Labs' Toronto Waterfront Tech Hub Must Respect Privacy, Democracy*, Star (Jan. 12, 2018), https://www.thestar.com/opinion/contributors/2018/01/12/sidewalk-labs-toronto-waterfront-tech-hub-must-respect-privacy-democracy.html; Bianca Wylie, *Sidewalk Toronto: Time to Take Data Governance Away from Sidewalk Labs *and* Waterfront Toronto*, Medium (Nov. 12, 2018), https://medium.com/@biancawylie/sidewalk-toronto-time-to-take-data-governance-away-from-sidewalk-labs-and-waterfront-toronto-cf6325b32cc7. Many of these issues have been raised at Waterfront Toronto's consultations through its Civic Labs initiative, which one of the authors has participated in. *See* Stephanie Chow, *Smart Cities. Smart Governance. Civic Labs on Digital Governance and Intellectual Property at Quayside*, WATERFRONToronto (Nov. 19, 2018), http://blog.waterfrontoronto.ca/nbe/portal/wt/home/blog-home/posts/Civic-Labs-at-Quayside-November-2018 (describing the Civic Labs meeting as integral in "shap[ing] the direction of the Sidewalk Toronto project"). For concerns with the project that go beyond data governance, see Nabeel Ahmed & Mariana Valverde, *The Waterfront Toronto Crisis: What Are the Options?*, Ctr. for Free Expression, https://cfe.ryerson.ca/key-resources/commentary/waterfront-toronto-crisis-what-are-options (last visited Jan. 14, 2019), which raises issues of public procurement and real estate development, among others.

[3] *See generally* Sidewalk Labs, Digital Governance Proposal for DSAP Consultation (2018), https://waterfrontoronto.ca/nbe/wcm/connect/waterfront/41979265-8044-442a-9351-e28ef6c76d70/18.10.15_SWT_Draft+Proposals+Regarding‡ata+Use+and+Governance.pdf?MOD=AJPERES [hereinafter Digital Governance Proposal]. Note that "urban data" is very generally defined as "data collected in the physical environment." *Id.* at 14.

[4] Statistics Canada developed its Research Data Centres (RDC) Program to respond to this tension. Some research can be conducted adequately using Statistics Canada Public

Releasing the data to others also undermines efforts to control against forms of misuse of the data, whether deliberate or unintentional. Ideally, the risks of re-identification and misuse could be mitigated through a mechanism through which data computations are performed in a secure and privacy-protective manner without releasing the raw data, and where all data sharing is transparent and auditable. Transparency does not mean that all data sharing becomes a matter of public view, but rather that data-sharing activities are made visible to organizations and regulators in appropriate circumstances, recognizing the potential confidentiality interests. We propose such a mechanism: a "safe sharing site."

Data sharing is a key activity within the emerging data ecosystem, not just in relation to smart city projects. It is an activity that sits at the crossroads of privacy concerns and the broader challenges of data governance. A safe sharing site functions as a piece of what we call "legal-digital infrastructure"—infrastructure that goes beyond solving the technical issues associated with data sharing and ensures that it can work with *many* different forms of legal regulation and governance.[5] As we outline in this Article, the design we envisage involves a legal interface that allows the safe sharing site to work with a variety of legal contexts. These can include privacy law (including those of multiple jurisdictions) as well as new forms of governance meant to address broader questions about social surveillance, algorithmic accountability, or data monopolies. A safe sharing site does not solve these problems, but rather creates an infrastructure for solutions by enabling forms of visibility and auditability of data use. In this sense, a safe sharing site supports multiple use contexts and multiple forms of data governance.

Our argument for safe sharing sites proceeds as follows. Part I discusses data sharing as an important activity within the data economy and describes its corresponding risks. We outline why current solutions such as "open data," which emphasize de-identification of personally identifiable information, as well as contractual safeguards, are flawed. Part II outlines the safe sharing site as our alternative means of managing these risks and as a way to provide an infrastructure that can support different forms of legal regulation and governance. Part III discusses data sharing in three different use cases

Use Microdata Files, which are de-identified datasets. However, some forms of research require access to more accurate individual level data. Statistics Canada allows this through the RDC Program, which involves accessing the data within a security facility. *See The Research Data Centres (RDC) Program*, STAT. CAN., https://www.statcan.gc.ca/eng/rdc/index (last visited Apr. 14, 2019).

[5] *See infra* Part IV for an outline of the various usages of the term "data trust."

that all raise privacy concerns in different legal contexts: data protection law, access by law enforcement, and the litigation process. We show how safe sharing sites can address data-sharing risks across these different contexts, illustrating the modular nature of the safe sharing site solution.[6] In Part IV, we show how safe sharing sites can work in relation to data governance concerns and models. For example, while a safe sharing site is different from a "data trust"—a governance mechanism for making decisions about some categories of data—safe sharing sites can work alongside such trusts. We also outline a number of issues regarding how safe sharing sites themselves should be governed.

I
Data Sharing and Data-Sharing Risks

A. Why Sharing?

Organizations in both the private and public sector increasingly see data sharing as key to the success of data-driven innovation.[7] For example, a recent United Kingdom report argues that "[t]o continue developing and applying AI, the UK will need to increase ease of access to data in a wider range of sectors."[8] Reducing the complexities involved in data sharing is therefore seen as central to a successful data economy. This raises the question of how to enable data sharing in a manner that still enables robust protection of privacy and other rights and interests.

One aspect of providing a solution that reduces these complexities, in our view, is to focus on data sharing as a distinct data activity

[6] Modularity in software programming is a technique to manage complexity. Software modules are created to have specific functions and then interface with other modules. *See generally Importance of Modularity in Programming*, Aspect-Oriented Software Dev. (Jan. 18, 2018), http://aosd.net/importance-of-modularity-in-programming (explaining how modularity helps manage complexity).

[7] "Data sharing" can be broadly understood to incorporate all models of access to and transfer of data between organizations, although some scholars narrow this term to exclude contexts where data protection laws apply. *See, e.g.*, Heiko Richter & Peter R. Slowinski, *The Data Sharing Economy: On the Emergence of New Intermediaries*, 50 Int'l Rev. Intell. Prop. & Competition L. 4 (2019) (giving data sharing a "more specific" definition in the regulatory context). We use it here in the broad sense and include contexts where data protection laws (or other privacy laws) apply.

[8] Wendy Hall & Jérôme Pesenti, Growing the Artificial Intelligence Industry in the UK 2 (2017), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf. On the importance of data to innovation generally, see *Data-Driven Innovation for Growth and Well-Being*, Org. for Econ. Co-operation & Dev., https://www.oe.cd/bigdata (last visited Apr. 14, 2019), which describes data-driven innovation as a "key pillar in 21st century sources of growth."

and address the issues associated with that activity in a modular fashion. This is in contrast to viewing data sharing as one aspect of a more general activity—like conducting research—and addressing data sharing within the context of that more general activity.[9] This view subsumes data sharing within the regulatory approaches to those distinct contexts, whereas our approach treats data sharing consistently across multiple contexts.

As Jack Balkin perceptively noted, the rise of new methods of data analytics over the last decade or more has shifted the focus of privacy concerns from data collection to data use.[10] But paying attention to data use rather than data sharing occludes the growing complexity of the data ecosystem and the multiple players involved. Many data-use discussions remain within a framework preoccupied with the individual-organization or individual-state relationship and the life cycle of data involved in that relationship, from collection to use. Data sharing asks us to also look more directly at relationships between organizations and the interests involved in those relationships.

Privacy is a central concern in data sharing. However, data sharing can take different forms and these different forms can raise distinct issues. For example, researchers often want access to data held by others to analyze it, but may only be interested in the aggregate results of this analysis. Other privacy debates concerning data-sharing practices have focused on data linkage between data sets containing personally identifiable information. With data linkage, an organization seeks to combine an existing data set with another data set (which may be controlled by another organization). Where this involves individual-level records, it allows an organization to get a more detailed picture of the individuals involved. This can raise important issues, including questions of consent and the ethics of profiling.[11]

Even where technology companies have their own existing data, they might want to provide assurances that they *cannot* link this data

[9] *See, e.g.*, Digital Economy Act 2017, c. 30, pt. 5, ch. 5 (Eng.), http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted (permitting accredited researchers to gain access to de-identified data for research purposes). This approach regulates data-sharing practices associated with research using public data but does not address other data-sharing contexts.

[10] Jack M. Balkin, *The Constitution in the National Surveillance State*, 93 MINN. L. REV. 1, 12 (2008) ("Government's most important technique of control is no longer watching or threatening to watch. It is analyzing and drawing connections between data.").

[11] This was what was at issue in a recent privacy scandal involving Statistics Canada. Statistics Canada wanted to collect financial information from financial institutions that it could link to demographic information that it already possessed. *See* Peter Zimonjic, *Privacy Commissioner Launches Probe into StatsCan over Collection of Financial Data*, CBC NEWS (Oct. 31, 2018, 1:35 PM), https://www.cbc.ca/news/politics/personal-financial-information-statistics-canada-1.4885945.

to other data sets already under their control. This was part of the recent controversy surrounding a London-based health AI firm, DeepMind, and its development of a health app involving patient data from the UK's National Institute of Health. DeepMind is owned by Alphabet, Google's parent company. When Google announced that it would take over this app to allow the product to scale, privacy advocates expressed concern regarding the possibility of data linkage between Google's data and the health information—something that was originally promised not to happen.[12]

Data sharing also implicates data governance concerns that go beyond individual privacy. Some emerging issues, such as algorithmic accountability and the ethics of profiling, are driven by practices that increasingly involve data sharing.[13] But data sharing can also potentially provide *solutions* to these concerns. For example, initial debates regarding algorithmic accountability emphasized algorithmic transparency, or greater openness regarding data practices.[14] Sometimes the problems of bias in algorithmic decisionmaking arise due to problems of bias in the data that is used to train the system, so the ability to access and review training data can be one strategy to address bias in forms of AI like machine learning.[15] More generally,

[12] *See* Alex Hern, *Google 'Betrays Patient Trust' with DeepMind Health Move*, GUARDIAN (Nov. 14, 2018, 7:14 AM), https://www.theguardian.com/technology/2018/nov/14/google-betrays-patient-trust-deepmind-healthcare-move (reporting that the app's cofounder stated that "at no stage [would] patient data ever be linked or associated with Google accounts, products or services"); *see also* Cara McGoogan, *NHS Illegally Handed Google Firm 1.6m Patient Records, UK Data Watchdog Finds*, TELEGRAPH (July 3, 2017, 3:46 PM), https://www.telegraph.co.uk/technology/2017/07/03/googles-deepmind-nhs-misused-patient-data-trial-watchdog-says (noting that DeepMind and the NHS were found to have violated the UK's Data Protection Act of 2017 through their information sharing arrangement, which did not provide enough transparency to patients or safeguards for the data).

[13] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (discussing algorithms' potential to commit employment discrimination); Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103 (2018) (finding that the government's use of predictive algorithms is insufficiently transparent); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017) (raising ethical challenges to society's increasing use of AI); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 18–24 (2014) (arguing for procedural safeguards in industries that use predictive algorithmic scoring); Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. 1735 (2015) (raising issues with blacklisting systems such as no-fly lists and no-vote lists based on algorithmic data).

[14] *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015) (arguing for more transparency, particularly in corporate and financial scoring systems).

[15] *See* AI NOW INST., ALGORITHMIC ACCOUNTABILITY POLICY TOOLKIT 28 (2018), https://ainowinstitute.org/aap-toolkit.pdf (defining training data as "[t]he input data used by a machine learning algorithm to find patterns"); PETRA MOLNAR & LEX GILL, THE

there are arguments that access to data can also be a solution to the growing concerns around data monopolies.[16] For example, a discussion paper from Canada's Competition Bureau recently stated that "[p]roviding access to the data may be an appropriate quasi-structural remedy allowing potential competitors in the downstream market to overcome their main barrier to entry."[17] Kelsey Finch and Omer Tene point out that municipalities like New York, San Francisco, and São Paulo "have revised rules or brought bills requiring Uber to share granular data about individual trips" for various regulatory purposes.[18] In the public sector, most liberal democracies have both privacy laws *and* freedom of information laws to govern their information practices.[19] While privacy laws have proliferated globally in relation to private sector data practices, there are no parallel freedom of information laws governing the private sector. Private sector laws providing access to data, consistent with competition law concerns, could be the next frontier in data governance.

The safe sharing site solution offered in this Article focuses on data sharing as a distinct activity and seeks to reduce key data risks associated with that activity. However, it does so on the assumption that data sharing will occur across a wide range of contexts that raise

CITIZEN LAB & INT'L HUMAN RIGHTS PROGRAM, BOTS AT THE GATE: A HUMAN RIGHTS ANALYSIS OF AUTOMATED DECISION-MAKING IN CANADA'S IMMIGRATION AND REFUGEE SYSTEM 65 (2018), https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf (recommending Canada create an independent oversight body to audit source code and training data used in its immigration system).

[16] *See* HOUSE OF LORDS ARTIFICIAL INTELLIGENCE COMMITTEE, AI IN THE UK: READY, WILLING AND ABLE?, 2017–19, HL 100, ¶ 129 https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm ("The monopolisation of data demonstrates the need for strong ethical, data protection and competition frameworks in the UK, and for continued vigilance from the regulators. We urge the Government, and the Competition and Markets Authority, to review proactively the use and potential monopolisation of data by the big technology companies.").

[17] COMPETITION BUREAU CAN., BIG DATA AND INNOVATION: IMPLICATIONS FOR COMPETITION POLICY IN CANADA 28 (2017), http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/vwapj/Big-Data-e.pdf/$file/Big-Data-e.pdf; *see also* Teresa Scassa, *Statistics Canada Faces Backlash over Collection of Personal Financial Information (or: Teaching an Old Law New Tricks)*, TERESA SCASSA BLOG (Oct. 31, 2018, 7:50 AM), http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=291:statistics-canada-faces-backlash-over-collection-of-personal-financial-information-or-teaching-an-old-law-new-tricks&Itemid=80 (commenting on Statistics Canada's attempt to seek access to financial data held by the private sector).

[18] Kelsey Finch & Omer Tene, *Smart Cities: Privacy, Transparency, and Community*, *in* CAMBRIDGE HANDBOOK OF CONSUMER PRIVACY 125, 136 (Evan Selinger et al. eds., 2018).

[19] For example, in the United States there are the federal Privacy Act, 5 U.S.C. § 552a (2012), and the Freedom of Information Act, 5 U.S.C. § 552 (2012). In Canada, there are the federal Privacy Act, R.S.C. 1985, c P-21, and the Access to Information Act, R.S.C. 1985, c A-1.

different normative questions and are subject to different regulatory regimes. Regulatory regimes may differ based on jurisdiction, the legal model employed, or the norms associated with the specific data practice. As we will outline, even though a safe sharing site addresses data sharing in a modular fashion in order to reduce complexities, it must also be able to interface effectively with multiple regulatory regimes.

B. PII as Gatekeeper

Data protection law, and its underlying Fair Information Practice Principles (FIPPs), is the twentieth century's major data governance paradigm.[20] The paradigm is focused on individual privacy, which is understood in terms of individual control over personal information. Although this paradigm is sometimes conflated with "notice-and-choice" models like the one that has developed in the United States under the Federal Trade Commission (FTC),[21] most articulations of FIPPs contemplate a broader idea of individual control than an initial take-it-or-leave-it choice and a broader range of obligations on data processors than consent. These obligations include, but are not limited to, data minimization, data accuracy, and data security.[22] The European Union's new General Data Protection Regulation (GDPR)

[20] *See* U.S. Dep't of Health, Educ. & Welfare, Sec'y's Advisory Comm. on Automated Pers. Data Sys., No. (OS)73-94, Records, Computers, and the Rights of Citizens (1973) (recommending development of a Fair Information Practice code and highlighting its importance); 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), 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR] (recognizing a person's right to privacy in their data as a fundamental right); Org. for Econ. Co-operation & Dev., Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data (1980) (outlining guidelines for effective protection of data and privacy).

[21] *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 Colum. L. Rev. 583, 594, 634–36 (2014) (describing the FTC's notice-and-choice approach). The FTC embraces a very narrow articulation of FIPPs (notice, choice, access, security, and enforcement) whereas many other jurisdictions—like Canada and the EU—follow the broader set of principles outlined in the OECD Guidelines that include collection limitation, data quality, use limitation, security safeguards, openness, individual participation, and accountability. Collection and use limitations contemplate the idea of data minimization, or the idea that an organization should not collect more data than is necessary. *See* Org. for Econ. Co-operation & Dev., Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2013).

[22] *See* Michael Birnhack, *A Process-Based Approach to Informational Privacy and the Case of Big Medical Data*, 20 Theoretical Inquiries L. 257, 265 (2019).

goes even further and adds additional individual entitlements such as data portability and some forms of algorithmic transparency.[23]

The point of entry for the different regulations modelled on FIPPs is Personally Identifiable Information (PII): When an organization deals with PII, then FIPPs apply, and if an organization does not deal with PII, then FIPPs do not apply.[24] Because PII is the gatekeeper for most data protection law, the easiest way to enable broad data sharing is to de-identify the data at issue; once stripped of PII, the basic argument goes, the data is no longer subject to regulation. If the data is no longer subject to regulation, then many of the legal risks and complexities associated with data sharing vanish.

De-identification is one of the core strategies in the Sidewalk Toronto project. To deal with privacy questions—including consent for the collection of data in civic spaces—the project proposes to de-identify the data at the point of collection, which takes it out of the regulatory framework. Because the data will be de-identified, the project proposes that the data can be freely and publicly shared in order to spur urban innovation.[25] In general, the open data movement more broadly relies upon this strategy for addressing potential privacy problems associated with the public release of data.[26]

[23] *See* GDPR, *supra* note 20, at arts. 5, 20, 22. Data portability refers to the idea that users should be able to receive their own personal information in a "structured, commonly used and machine-readable format" that allows them to take it to another organization. *Id.* at art. 20. Algorithmic transparency refers to the right, in some circumstances, to an explanation of the basis for an automated decision. *See* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 192 (2019) (explaining the GDPR's algorithmic accountability principles); Gabriela Zanfir, *The Right to Data Portability in the Context of the EU Data Protection Reform*, 2 INT'L DATA PRIVACY L. 149, 157 (2012).

[24] For example, Canada's federal private sector data protection law, Personal Information Protection and Electronic Documents Act (PIPEDA), S.C. 2000, c 5, § 4(1), only applies to organizations that collect, use, and disclose "personal information." Similarly, the GDPR applies to the processing of "personal data." *See* GDPR, *supra* note 20, at art. 2. If organizations are not dealing with data that can directly or indirectly identify an individual, then the regulations do not apply.

[25] *See* DIGITAL GOVERNANCE PROPOSAL, *supra* note 3. The proposal also offers hypothetical examples of the types of data it will collect: pedestrian count to manage traffic, video cameras to capture park usage, and energy and environmental condition trackers to measure usage. *Id.* at 25, 28.

[26] *See, e.g.*, FUTURE OF PRIVACY FORUM, ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY, NOTHING TO HIDE: TOOLS FOR TALKING (AND LISTENING) ABOUT DATA PRIVACY FOR INTEGRATED DATA SYSTEMS 15 (2018), https://fpf.org/wp-content/uploads/2018/09/FPF-AISP_Nothing-to-Hide.pdf (including de-identification in a list of "key privacy tools"); Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1967, 1981 (2015) (explaining that the Privacy Act prohibits federal agencies from releasing identifiable information gathered during data collection); Frederick Zuiderveen Borgesius et al., *Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework*, 30 BERKELEY TECH. L.J. 2073, 2116–18 (2015) (noting that computer scientists and other experts advocate for data

This is a flawed strategy. First, the boundary between PII and non-PII is unstable and there is always a risk of re-identification. Strategies to mitigate these risks affect the accuracy of the data. If the goal of data sharing is to spur innovation and contemporary computational methods require large and accurate datasets, then there is a fundamental tension between managing the risk of re-identification and ensuring sufficient data accuracy. Second, privacy—as understood in terms of individual control over personal information—is not the only normative issue raised by data sharing. Data about persons, even if not individually identifiable and even if aggregated, raises numerous social, political, and economic questions.[27] Therefore, even if data sharing does not involve PII, it does not take place in a norm-free zone. We elaborate on these points below and argue that alternative strategies for data sharing will have to incorporate ideas of transparency and auditability of data use if they are to enable responsible risk management and data accountability.

C. Data Accuracy and Re-Identification Risks

A considerable body of research shows that "anonymous" data can be re-identified in a variety of ways.[28] The frailty of anonymization is gradually entering the public consciousness through a number of high-profile examples of individuals being re-identified from publicly released de-identified data sets. Researchers have succeeded in re-identification attacks on publicly released "anonymous" Netflix

anonymization; anonymized data is de-identified, which allows for a balance of data and privacy interests).

[27] See *infra* Section I.D for a discussion of norms in the data-sharing context. Sidewalk Labs has also recently released a more detailed Master Innovation and Development Plan; the Plan discusses using environmental sensors and creating a real-time map of open spaces. *See* Sidewalk Labs, Toronto Tomorrow: A New Approach for Inclusive Growth 16, 176–86 (2019), https://quaysideto.ca/wp-content/uploads/2019/07/MIDP-Volume-2-Printer-Friendly.pdf.

[28] For example, see the pioneering work of Latanya Sweeney. Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* (Carnegie Mellon Univ., Data Privacy Working Paper No. 3, 2000), which concludes that, even with a few data points, it can be easy to identify a person through de-identified data. For a discussion of this work in a legal context, see generally Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010), which describes re-identification techniques and the resulting disruption of privacy law. *See also* Lisa M. Austin, *Reviewing PIPEDA: Control, Privacy and the Limits of Fair Information Practices*, 44 Can. Bus. L.J. 21, 35–36 (2006) (discussing the consequences of re-identification under Canada's PIPEDA); Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 Wash. L. Rev. 703, 711–14 (2016) (presenting the re-identification risks associated with "quasi-identifiers" such as age and gender); Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. Rev. 1814, 1836–48 (2011) (noting traceability of IP addresses and discussing re-identification of Google, AOL, and Netflix users).

user data,[29] AOL data,[30] and Australian health data.[31] Other research suggests that mobile usage data is easily re-identified,[32] which is of particular concern in the smart city context. It is increasingly clear that complete anonymity is rarely technologically possible and that instead, we must focus on the varying degrees of risk of re-identification.

There are a number of developed techniques that can mitigate the risk of re-identification, but they all involve trade-offs between the protection of privacy and the resulting utility of the data. For example, *k*-anonymity[33] is based on the principle that privacy is achieved if an individual cannot be distinguished from $k - 1$ other individuals in a publicly released dataset. Think of a dataset as a series of rows and columns. Each row represents a distinct individual and every column provides specific characteristics of that individual (such as gender, age, or health status). Even when explicit identifiers like "name" are removed, combinations of characteristics can uniquely identify individuals (for example, if there is only one row with a female of twenty-eight years of age with a broken leg). A data set with $k = 3$ means that for every set of identifying characteristics, there are at least three individual rows in the dataset with those characteristics (for example, three females of twenty-eight years of age with a broken leg). The higher the value of *k*, the lower the risk of re-identification.

Broadly, *k*-anonymity can be achieved by removing information about individuals from a dataset or by broadening the granularity at which information is given. For example, a dataset consisting of the age and gender of a set of individuals may enable re-identification if each combination of age and gender is unique. To avoid re-identification, one could remove age, thus allowing individuals to blend into a crowd with *k* individuals of the same gender, or one may

[29] Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, PROC. 2008 INST. ELECTRICAL & ELECTRONICS ENGINEERS SYMP. ON SECURITY & PRIVACY 111 (demonstrating the de-anonymization of Netflix data with little auxiliary information).

[30] *See* Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), https://www.nytimes.com/2006/08/09/technology/09aol.html.

[31] CHRIS CULNANE ET AL., HEALTH DATA IN AN OPEN WORLD: A REPORT ON RE-IDENTIFYING PATIENTS IN THE MBS/PBS DATASET AND THE IMPLICATIONS FOR FUTURE RELEASES OF AUSTRALIAN GOVERNMENT DATA (2017), https://www.researchgate.net/publication/321873477_Health_Data_in_an_Open_World.

[32] *See* Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 SCI. REP. 1376 (2013) (using cellular location data to determine that ninety-five percent of the individuals studied could be uniquely identified in the dataset).

[33] Latanya Sweeney, k-*Anonymity: A Model for Protecting Privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYS. 557 (2002).

choose to broaden the age category by grouping individuals into an age range (e.g. 20–29, 30–39). *K*-anonymity protects privacy but destroys the ability to query for phenomena whose frequency occurs below the *k*-threshold. In addition, increasing the number of categories of data associated with each individual will result in greater loss of detail, as individuals must be identical to $k - 1$ other individuals across all categories of data.[34] Further, the values in said columns cannot be exactly the same for a particular group in order to avoid vulnerability to a homogeneity attack in which it is sufficient to find which group a particular individual belongs to.[35]

Another technique is differential privacy, which measures the amount of privacy lost by individuals in a dataset.[36] Differential privacy assigns a numerical value to this privacy loss, which allows for an approach that explicitly measures the trade-off of privacy loss with the accuracy of data in a dataset. The amount of privacy loss is defined as the probability that an adversary could make two queries on the dataset whose results differ only by a single individual. For example, suppose a database contains information about individuals with a particular virus but forbids queries that would return a single individual (for example, does Person X have HIV?). An attacker can still learn this information by making the related queries: How many individuals have HIV? And how many individuals who are not named X have HIV? If the two queries differ only by one, then the attacker will have learned the information that was forbidden.

Unlike *k*-anonymity, differential privacy is based on probabilities, so it uses a different mechanism to hide the true value of the data to preserve privacy—the introduction of noise or false data. By altering the data of random individuals in a dataset, differential privacy introduces uncertainty regarding the correctness of the data. This decreases the probability that an adversary can be sure that two queries truly differ by a single individual. This method of introducing noise is often called "randomized response," as it is implemented by randomly changing the responses of individuals to the survey that would form the dataset. The disadvantages of differential privacy are similar to those of *k*-anonymity—to achieve a sufficient level of pri-

[34] *See* Charu C. Aggarwal, *On* k*-Anonymity and the Curse of Dimensionality*, PROC. 31ST INT'L CONF. ON VERY LARGE DATA BASES 901, 906–09 (2005), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.3155&rep=rep1&type=pdf (noting that data becomes less precise with increased dimensionality).

[35] Debasis Mohapatra & Manas R. Patra, *Analysis of* k*-Anonymity for Homogeneity Attack*, 3 INT'L J. ADVANCES COMPUTER SCI. TECH. 30 (2014) (exploring homogeneity attacks' effects when database size and *k*-value vary).

[36] *See* Cynthia Dwork, *Differential Privacy*, 33 INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES & PROGRAMMING, PART II, 1, 8–11 (2006) (defining differential privacy).

vacy, a certain amount of noise must be added. Adding noise is equivalent to intentionally adding errors to the dataset. This can lead to some erroneous conclusions being drawn from analysis of the data.

Open data exacerbates the difficulties of de-identification. Released publicly, the data can be used by anyone for any purpose and at any time in the future. Because there are measurable trade-offs between privacy and utility of the data, it is not clear that adequately addressing the privacy risks associated with open data proposals will result in data that will be of sufficient utility for the many different uses contemplated—as well as future uses that are presently unknown. This is particularly true in the context of applications that require large datasets.[37] In addition, it is difficult to properly measure the risk of re-identification for publicly released data in an era of rapidly-increasing data collection and analysis: We just do not know what kind of information or advanced data analytics will be available in five to ten years that could be used to re-identify the data.

Where the model for thinking about access to data is the release of de-identified data sets, the accompanying policy solutions for addressing the privacy-accuracy trade-off will not necessarily favor privacy concerns in all instances. Instead, these solutions will likely take some form of risk/benefit analysis. For example, Finch and Tene propose doing both a Privacy Impact Analysis (PIA) and a Data Benefit Analysis (DBA) in order to "put[ ] a project's benefits and risks on an equal footing."[38] Finding a better technical solution to the PII concern will reduce the potential number of cases where privacy is compromised in order to achieve a benefit and reduce some of the complexities involved in the decisionmaking.

D. There Is No Norm-Free Data Sharing

Even if de-identification were technically possible, the strategy does not address the many potential normative considerations that accompany data sharing. As Helen Nissenbaum has argued about the

[37] *See* Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, PROC. 14TH ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 70 (2008) (evaluating the utility-privacy tradeoff when data is sanitized).

[38] Finch & Tene, *supra* note 18, at 131; *see also* FUTURE OF PRIVACY FORUM, CITY OF SEATTLE OPEN DATA RISK ASSESSMENT FINAL REPORT 13 (2018), https://fpf.org/wp-content/uploads/2018/01/FPF-Open-Data-Risk-Assessment-for-City-of-Seattle.pdf (presenting a model benefit-risk analysis for open data); JULES POLONETSKY ET AL., FUTURE OF PRIVACY FORUM, BENEFIT-RISK ANALYSIS FOR BIG DATA PROJECTS 1 (2014), https://fpf.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf (noting that organizations use privacy impact assessments to determine privacy risks with a new project and balance those risks with the product's benefits).

public/private divide in privacy debates, there is no "public" space that involves the absence of norms regarding the appropriate flow of information.[39] Similarly, we should not think that the PII/non-PII divide somehow only involves norms on the PII side. Sharing data about people, even if people are not personally identifiable, still involves norms of appropriate and legitimate information flows. As we outline below, some of these norms are related to traditional privacy concerns; others point to a broader constellation of data governance concerns. In other words, even if there are no appreciable risks of re-identification in a data set, there might still be other data risks that need to be managed or regulated.

One area where the PII/non-PII distinction breaks down—even in relation to concerns that can be classified as privacy issues—is in the methods of online surveillance associated with behavioral advertising and web analytics. For example, Google explicitly interprets PII as used within its own contracts and policies to *exclude* cookies and IP addresses whereas the Office of the Privacy Commissioner of Canada has indicated that an IP address and cookies can be personal information, although context is important.[40] The FTC also points to the privacy concerns associated with non-PII, and the agency applies its principles to data that "reasonably could be associated with a particular consumer or with a particular computer or device."[41] Collecting individual-level data, and using it in various forms of profiling, impacts how websites and advertisers interact with individual users. This kind of surveillance can impact individuals even when the practice may not formally involve PII. In many ways, it points to the folly of delineating a category of information (non-PII) that can be assigned a stable privacy risk (low); in reality, the risk level should be assessed within the context of the data use—especially when those uses involve complex and opaque data analyses.

[39] Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 137 (2004) ("A central tenet of contextual integrity is that there are no arenas of life *not* governed by *norms of information flow*, no information or spheres of life for which 'anything goes.'").

[40] *Compare Understanding PII in Google's Contracts and Policies*, GOOGLE, https://support.google.com/analytics/answer/7686480?hl=EN (last visited Apr. 14, 2019), *with* OFFICE OF THE PRIVACY COMM'R OF CAN., REPORT ON THE 2010 OFFICE OF THE PRIVACY COMMISSIONER OF CANADA'S CONSULTATIONS ON ONLINE TRACKING, PROFILING AND TARGETING, AND CLOUD COMPUTING (2010), https://www.priv.gc.ca/media/1961/report_201105_e.pdf.

[41] FED. TRADE COMM'N, FTC STAFF REPORT: SELF-REGULATORY PRINCIPLES FOR ONLINE BEHAVIORAL ADVERTISING 25 (2009), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-behavioral-advertising/p085400behavadreport.pdf.

In fact, there is a critical literature on data that highlights many of the normative dimensions of data sharing and use that go beyond privacy, but which have not yet become mainstream in policy discussions.[42] Consider the lessons that have come out of creating content management systems—software that facilitates organizing digital content—for Indigenous cultural heritage materials. As Kimberly Christen describes in relation to the development of the Mukurtu CMS:

> Instead of assuming that information *wanted to or should be* open, free, and available to "anyone with an Internet connection," our development process emphasized the underlying *sociality of information* and its reliance on, and embeddedness within, ethical systems of relation and action in which people negotiate the creation, reproduction, and distribution of knowledge based on multiple and interrelated factors and situations.[43]

This approach is sensitive to the political aspects of data. In the Indigenous context, this includes the backdrop of colonialism and the many questions it raises regarding where heritage materials are stored, how they were collected, by whom, and what community protocols should animate access norms. For example, Christen explains how protocols were developed that tracked "family and place-based relations, followed by community status defined by peoples' relations to both one another and traditional community knowledge" in a manner that mapped "*preexisting* social norms concerning the creation, reproduction, and distribution of knowledge within the community."[44] These issues are being raised on a much larger scale in the growing movement for "indigenous data sovereignty."[45]

Data circulates within a social context and simple calls for "open access" to de-identified data risk erasing the political, social, economic, and ethical dimensions of data access. If data sharing is to be

[42] For example, Surveillance Studies as well as Science and Technology Studies both look at data related practices from a much broader analytic lens than privacy. For a good overview, see David Murakami Wood, *Situating Surveillance Studies*, 19 SURVEILLANCE & SOC. 52 (2009), which analyzes the normative roles of surveillance as a means of control or empowerment of the citizenry.

[43] Kimberly Christen, *Does Information Really Want to Be Free?: Indigenous Knowledge Systems and the Question of Openness*, 6 INT'L J. COMM. 2870, 2887 (2012).

[44] *Id.* at 2885.

[45] *See generally* INDIGENOUS DATA SOVEREIGNTY: TOWARD AN AGENDA (Tahu Kukutai & John Taylor eds., 2016) (discussing the need for data on indigenous people as well as the need for indigenous people to have control and sovereignty over that data); *The First Nations Principles of OCAP®*, FIRST NATIONS INFO. GOVERNANCE COMM., https://fnigc.ca/ocapr.html (last visited Apr. 27, 2019) ("OCAP® ensures that First Nations own their information and respects the fact that they are stewards of their information, much in the same way that they are stewards over their own lands.").

safely facilitated, then it is important to understand that there is a broad array of interests at stake that go beyond the risk of re-identification. Therefore, even if de-identified data is released to an organization, there is still the need for methods to manage those additional risks. As we discuss in the following section, relying upon contractual arrangements or social norms—like reputational risks—is insufficient. There needs to be an infrastructure that supports transparency and auditability.

E. Data Sharing and Data Accountability

We can think of open data as one response to certain desired specifications in relation to data sharing. Those specifications are: (a) facilitating the sharing of data for innovation purposes but also in response to other data governance concerns including those regarding data monopolies, while (b) protecting against problematic uses including, but not limited to, those that violate privacy. The "open data" response to these specifications is to rely upon data de-identification to address concerns about problematic uses. This is a frail solution due to the issues raised in the previous sections, including the impacts of re-identification risk mitigation strategies on data accuracy and the need to think about privacy and other normative concerns more broadly than a narrow focus on PII permits.

Before outlining our alternative proposal for implementing these data-sharing specifications, we describe one more concern: the need for transparency and auditability in relation to data uses. Once one accepts the proposition that there is no such thing as a norm-free sharing space occupied by "anonymous" data, what is left is a variety of data uses and risks that must be regulated and governed. While we will later show that this involves multiple legal considerations, including but not limited to data protection laws, the basic foundation for these different aspects of data accountability is transparency and auditability.

The recent data-sharing scandals plaguing Facebook highlight the fragility of current accountability for data sharing. In December 2018, the *New York Times* reported that Facebook had been providing several large technology companies such as Microsoft, Netflix, and Spotify with access to Facebook users' personal data.[46] One notable aspect of the controversy remains whether the data sharing violated a 2011 consent decree with the FTC that requires Facebook to obtain

[46] *See* Gabriel J.X. Dance et al., *As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants*, N.Y. Times (Dec. 18, 2018), https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html.

consent for data sharing. Facebook has argued that this sharing falls under a consent exemption for "service providers" whom Facebook considers extensions of itself.[47] It is far from clear that Facebook is correct in its interpretation of "service providers"—the kinds of relationships disclosed by the *New York Times* are data *partnerships* where both parties obtain benefits from the sharing of data rather than a relationship where one party carries out activities on behalf of another and under their direction.

The "service provider" exemption is similar to the distinction in the GDPR between "data controllers" and "data processors." The latter process data "on behalf of the controller" and their identities do not have to be disclosed to data subjects in order to obtain valid consent—even under the GDPR's strong opt-in consent provisions.[48] However, under the GDPR, the identities of these data processors *do* have to be disclosed as a matter of the general transparency provisions, which are broader than the transparency requirements associated with informed consent.[49] The general trend in data protection law, therefore, is towards transparency of the parties involved in the data ecosystem quite apart from consent considerations. Because under the GDPR data processors can have direct compliance obligations, this transparency is important for facilitating accountability.[50] Therefore, even if Facebook is correct that its partners were processing data under its direction, and they therefore did not need consent, privacy policy disclosures of data sharing described in vague terms with unspecified partners are problematic under the GDPR model.

Facebook also claimed that its partners were required to respect Facebook users' privacy settings and were under contractual obligations to follow Facebook policies.[51] This is similar to the claims that Facebook made in relation to the Cambridge Analytica scandal. In that case, Aleksandr Kogan and Global Science Research collected data on as many as eighty-seven million Facebook users through an app called "thisisyourdigitallife" that collected data on both app users

[47] *See id.*

[48] Although not necessary for valid consent, the Working Party notes that "to comply with Articles 13 and 14 of the GDPR, controllers will need to provide a full list of recipients or categories of recipients including processors." *See* Article 29 Working Party, *Guidelines on Consent Under Regulation 2016/679*, § 3.3.1, WP259 rev. 01 (Nov. 28, 2017).

[49] *See* GDPR, *supra* note 20, at arts. 13–14; *see also* Article 29 Working Party, *supra* note 48, § 3.3.1.

[50] *See* GDPR, *supra* note 20, at rec. 22, art. 3(1).

[51] Dance et al., *supra* note 46.

and app users' Facebook friends.[52] The data collected was shared with Cambridge Analytica, a political consulting firm, and used to profile American voters for conservative political ad campaigns.[53] Facebook claimed that this was not a data breach because the data was collected according to the privacy settings that the app user and the user's friend had selected; the problem of inappropriate data access and use lay with Kogan's violation of Facebook's Platform policies that placed restrictions on this data use by app developers, including that the data not be passed on to third parties.[54] The UK's Information Commissioner's Office imposed a £500,000 fine on Facebook for both giving app developers access to user information "without sufficiently clear and informed consent" and for failing to secure the data in part because "it failed to make suitable checks on apps and developers using its platforms."[55]

The Facebook data-sharing stories emphasize the problems with establishing meaningful consent for data sharing in a complex data ecosystem, but they also highlight the need for transparency and auditability. Data protection law is moving towards requiring increased transparency in the context of data sharing, in addition to consent requirements.[56] The importance of transparency also goes beyond the data protection law framework: Data practices that are not visible in some manner are difficult to regulate. If practices are opaque to consumers or to regulators, then it is difficult for them to understand when these practices breach norms and require regulatory intervention. This is not an argument that transparency can substitute for other forms of regulation, but rather that transparency is an important foundation for those forms of (effective) regulation.[57]

Auditability is another important foundation for effective regulation. Contract can play a role in auditability, but it has many limits in

[52] *See* U.K. Info. Comm'r's Office, Investigation into the Use of Data Analytics in Political Campaigns: A Report to Parliament 38–39 (2018), https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf (discussing the Cambridge Analytica scandal).

[53] *See* Matthew Rosenberg et al., *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. Times (Mar. 17, 2018), https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

[54] Paul Grewal, *Suspending Cambridge Analytica and SCL Group from Facebook*, Facebook Newsroom (Mar. 16, 2018), https://newsroom.fb.com/news/2018/03/suspending-cambridge-analytica.

[55] *See* U.K. Info. Comm'r's Office, *supra* note 52, at 38.

[56] *See supra* notes 49–50 and accompanying text (discussing the requirements of the GDPR).

[57] For a general discussion of the research showing the failures of transparency as a regulatory strategy in relation to online contracting, see Florencia Marotta-Wurgler, *Even* More Than You Wanted to Know *About the Failures of Disclosure*, 11 Jerusalem Rev. Legal Stud. 63 (2015).

the new data ecosystem. For example, Facebook had many policies that were part of the contractual arrangements between the parties that set out their obligations in relation to the data shared.[58] Through these contractual arrangements, Facebook reserved the right to audit apps for compliance with their policies.[59] However, at least the UK Information Commissioner has found this arrangement inadequate.[60] Contract is a clumsy tool in the digital ecosystem, like using duct tape to hold together an arrangement because that is all you have, and it will work for a time. What is also needed are robust technical controls to manage sharing and enable transparency and auditability—not just for the organizations involved in the sharing, but also for the regulators who are charged with various forms of oversight.

II
The Safe Sharing Sites Solution

Across a wide variety of contexts, organizations need to be able to share data with other organizations.[61] However, as discussed, this sharing needs to occur in a manner that does not compromise the privacy or security of the data subjects or raise concerns regarding other types of problematic uses. Public release of de-identified data is a flawed solution because of the risks of re-identification, the reduction in data accuracy associated with risk mitigation techniques, and ongoing vulnerability to other forms of data misuse that go beyond individual privacy. Solutions that rely upon contractual arrangements between sharing partners to manage these risks are also flawed if compliance cannot be independently verified and if the contractual practices remain opaque to data subjects and regulators.

Safe sharing sites offer a different kind of solution. Similar to the basic de-identification strategy, safe sharing sites permit data sharing without giving access to PII. However, the manner in which they do so is different. Through a safe sharing site, a party holding raw data with PII could allow another party to analyze the data in select ways, while blocking them from viewing the raw data itself. For example, Organization *A* could allow Organization *B* to analyze its data

[58] *See* U.K. Info. Comm'r's Office, *supra* note 52, at 27.

[59] *See, e.g.*, *Statement of Rights and Responsibilities*, Facebook (Dec. 11, 2012), http://www.facebook.com/legal/terms [https://perma.cc/B8FK-NTQG] ("To ensure your application is safe for users, we can audit it.").

[60] *See id.*

[61] This is key to innovation, but access to data can also be important as an aspect of data governance, as already discussed. *See supra* Section I.A (discussing how data sharing is key to the success of data-driven organizations but also requires exploring systems of data governance).

according to *B*'s needs without actually disclosing its data to *B*. The safe sharing site provides a controlled environment where computations can be done on data and only the results of those computations leave the safe sharing site. These operations on the data would ideally be recorded and auditable. Important elements relevant to the legal regulation of data sharing could be made transparent to regulators through the creation of a registry of basic information regarding the claimed authority for processing the data, the jurisdiction of the data, and the use of the data. A safe sharing site is not a solution to all existing privacy questions (like consent) or emerging data governance concerns (like algorithmic accountability or data monopolies). Instead, it is meant to be a piece of infrastructure that permits an important data activity to occur under conditions that mitigate a number of important risks and make effective regulation possible across multiple legal contexts.

Some jurisdictions use secure centers to provide researchers with access to census data. The data that is released in aggregate public release files is usually not fine-grained enough for many types of social science research, so additional access to the underlying raw data is important for research. For example, Canada has Research Data Centres operated by Statistics Canada.[62] Researchers can get secure access if their research project is approved and they are willing to become "deemed employees" under the Statistics Act, therefore subject to that Act's rules and penalties regarding confidentiality.[63] If researchers do not want to submit to this process, they can indirectly access data sets through the "Real Time Remote Access" (RTRA) system.[64] On this latter model, researchers can query the data sets without getting full access to them. There are limits to this access. For example, the number of queries is limited, queries must be in the SAS language for statistical analysis,[65] and not all abilities are enabled because of the concern that some types of queries can leak data. Other jurisdictions also make de-identified public data available to researchers in secure environments, such as the UK model introduced in the recent Digital Economy Act.[66]

[62] *The Research Data Centres (RDC) Program*, *supra* note 4 (providing researchers with access, in a secure setting, to microdata).

[63] *See id.*; *see also* Statistics Act, R.S.C. 1985, c S-19, §§ 5(3), 6(1), 17(1), 30 (laying out some of the rules and penalties for using Statistics Canada's data).

[64] *The Real Time Remote Access (RTRA) System*, STAT. CAN., https://www.statcan.gc.ca/eng/rtra/rtra (last visited Apr. 24, 2019).

[65] *System Limitations*, STAT. CAN., https://www.statcan.gc.ca/eng/rtra/limitation (last visited Apr. 29, 2019).

[66] *See* Digital Economy Act 2017, c. 30, pt. 5, ch. 5 (Eng.), http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted (requiring, prior to disclosure, that information

Our safe sharing site proposal differs in a number of important respects from these other approaches being used in the research context. First, the Statistic Canada Data Centres and the RTRA system have been developed to meet the needs of academic researchers conducting statistical analysis of census data and there are many current limits to its access. We envision a solution that can meet the needs of multiple data-sharing contexts. One important technical concern that becomes more acute if this model is meant to apply to many data-sharing contexts is the ability to ensure that potentially harmful computations are restricted and that PII does not leak from the results of the computation. Second, the Data Centres exist to give access to data held by Statistics Canada under a particular set of rules (governed by the Statistics Act), and this determines how the Centres operate in terms of permissions and controls. In our safe sharing site model, the site would operate independently of the organizations who want to share the data. Therefore it is not that Organization *A* would create a safe sharing site in order to share data with Organization *B* but that *both* would enlist the services of a safe sharing site in order to facilitate their desired sharing of data. As we outline in the litigation example in Part IV, this independence can be important in a number of contexts. The UK model provides researchers with the actual raw data once it has been de-identified, whereas our model would provide researchers with the ability to do computations on the data without ever getting access to the raw data. This means that the accuracy level of the data available within a safe sharing site would be higher than in the UK model. Another important difference is that the UK model includes details regarding the accreditation of research whereas our model is agnostic in relation to the model of data governance adopted and can work with many different models.[67]

A safe sharing site is meant to be a modular solution that can work for many different parties in different contexts. We can find another source of inspiration in the software world. In a software module, the internal workings of the software remain opaque to users but, importantly, there are assurances that the modules do certain things and that they interface in specific ways with other things.[68] An

identifying "particular individual[s]" be processed such that "it is not reasonably likely that the person's identity will be deduced" even when "taken together with other information").

[67] *See infra* Part IV.

[68] Within software development, it is formally called "abstraction" when implementation is separated from the interface. *See, e.g.*, BARBARA LISKOV & JOHN GUTTAG, PROGRAM DEVELOPMENT IN JAVA: ABSTRACTION, SPECIFICATION, AND OBJECT ORIENTED DESIGN 2–12 (2001).

analogy is the experience of operating a car. Anyone who has operated one car can easily operate almost any other car because there is a consistent interface of steering wheel, pedals, turn signals, et cetera—even if different cars have vastly different and complex internal designs. Users can also operate a car despite knowing close to nothing about how it works internally. A safe sharing site works like a software module or a car in this way; it seeks to solve the problems associated with data sharing in a manner that can interface with multiple practical and legal contexts.

One of the challenges that our safe sharing site proposal faces is legal complexity. Legal complexity arises in a number of ways. For example, any proposed data sharing may involve both data subjects and organizations who are in multiple legal jurisdictions with various legal rules (existing and emerging) regulating the data. As we will show in our examples discussed in Part IV, even within a single jurisdiction, data sharing can occur within multiple legal contexts. Safe sharing sites need to be able to interact with these different legal regimes; they need a legal interface that allows different forms of legal regulation to interact with a safe sharing site.

This interface involves modes of transparency and assurances that enable a wide variety of accountability mechanisms. Visibility of sharing practices is an important component of effective legal regulation of any kind. A safe sharing site should include a registry of Organization *A*'s claimed lawful authority for sharing the data, as a mode of such transparency. The registry would make *A*'s claims transparent, but it would not necessarily certify the lawful authority as legitimate. However, having a registry like this would make it possible for various regulatory authorities to understand, and potentially investigate, some of *A*'s data practices. This registry could be made available to various regulators in a form that permits types of automated inquiries as well, enabling the development of new forms of regulatory tools.[69] Other elements of this registry would be information regarding the jurisdiction of the data and the uses of the data made by Organization *B*. The registry would be "public" in the sense that it is meant to enable forms of public regulation, but it need not be open to the public generally, as there may be confidentiality concerns associated with some data sharing. Rules regarding who can have access to this registry, or elements of it, would be up to each jurisdiction.

[69] *See* Lisa M. Austin, David Lie, Peter Yi Ping Sun, Robin Spillette, Mariana D'Angelo & Michelle Wong, Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project (June 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3203601 (discussing the need for new types of regulatory tools).

A legal interface capable of facilitating oversight and accountability should provide assurances and auditability—not just transparency. There has to be a way of ensuring that independent third parties should be able to come to the same conclusion about the activities that occur within the safe sharing site. Some of these might be assurances for activities that a safe sharing site is responsible for, such as the security of the data. But there may also be a need for independent verification of some aspect of the use of data for which the sharing organizations are responsible. This could include a full audit of the data processing, but it also could include a verifiable assurance that, for example, the data processing did not involve the processing of sensitive indicators.[70]

Other challenges concern the potential effects of implementing auditing requirements on privacy and other interests in the data. For example, Organization *B* may use proprietary methods to analyze data in a safe sharing site. If these methods were recorded, audits might reveal trade secrets or confidential information. Safe sharing sites do not themselves manage this tension between transparency and privacy; they provide an interface that allows for this tension to be managed by different legal regimes—both within a single jurisdiction and across multiple jurisdictions. This could be done by requiring auditable data operations and then leaving it to different substantive laws to set the rules for who has authority to audit and in what circumstances.

The safe sharing site proposal does not do away with the need for de-identification. This could still be an aspect of safeguarding the data while it is being analyzed. However, because the specific data use would be known, the de-identification could be tailored based on more precise knowledge of re-identification risks. For example, when applying differential privacy, the amount of noise added depends on the number of queries to the database. The amount of noise is increased as the number of queries grows.[71] If noise must be added before the number of queries is known, then a large amount of noise must be added to ensure that privacy guarantees are strong. However, if the computation is known beforehand, as it would be in a safe sharing site, noise can be added as needed and tailored to the specific computation. This provides a more accurate result without a negative impact on the desired level of privacy.

[70] See *infra* Section III.A for a proposed application of this approach to safeguarding personal information in online ad auctions.

[71] *See* Dwork, *supra* note 36, at 9–10.

Further, before Organization *B* is able to specify a computation to be performed on Organization *A*'s data, *B* would need to know certain facts about *A*'s data. A simple strawman could be used to provide a schema describing the types and number of data in *A*'s dataset. *B* could plan its computations based on this schema. Alternatively, *A* could provide *B* a suitably de-identified dataset that has been protected using *k*-anonymity or differential privacy. *B* could construct its computations using the de-identified dataset and then obtain the true and precise answer by performing the same computation on the raw dataset inside a safe sharing site. Without a safe sharing site, *B* would have to perform its computation on the de-identified dataset and achieve lower precision because of the privacy-protecting technology.

A similar technology is blockchain, which provides an immutable, sometimes publicly accessible, distributed ledger of events (such as payments).[72] While a blockchain is an effective technological mechanism for implementing accountability, it is often ineffective at providing privacy.[73] This is because all events recorded in the public chain are visible to all participants. In contrast, a safe sharing site would provide both auditability and privacy. Organization *A*'s data would not be disclosed and neither would the computations that Organization *B* performs on *A*'s data. Thus, a safe sharing site would provide benefits that blockchains are inherently unable to provide.

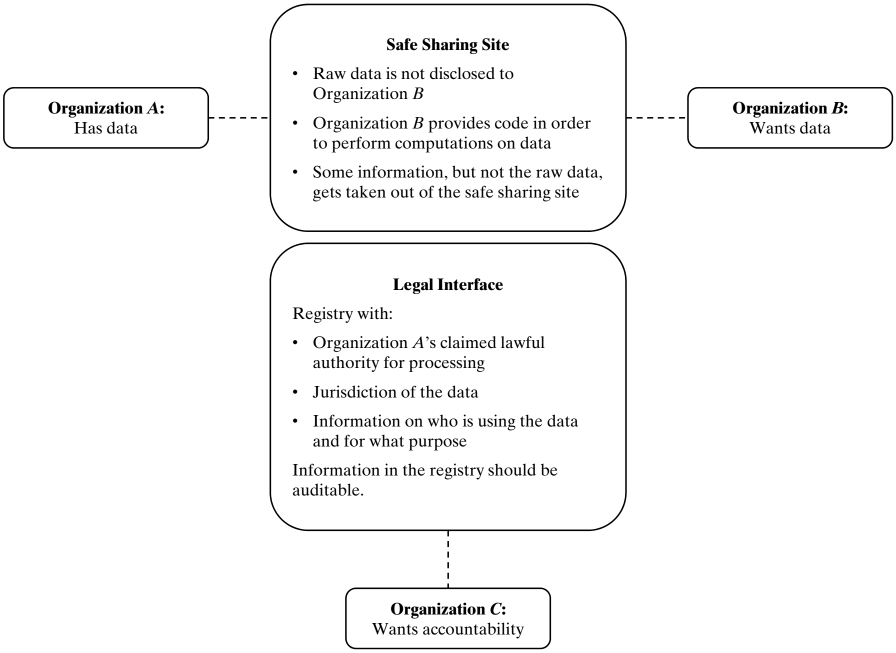
Finally, a related concern is whether safe sharing sites increase the consequences of a data breach. Like any entity that stores data, safe sharing sites might contain vulnerabilities or flaws that enable attackers to gain unauthorized access to raw and sensitive information. In this case, whether safe sharing sites increase the risk of a breach or increase the damage that a breach could cause depends on how the safe sharing site is used. On one hand, if each safe sharing site were only to contain data from a single organization, then the damage and risk would be similar to the organization itself being breached. In fact, if the safe sharing site is implemented by a competent party with sufficient oversight, the risks may in fact be lower than when an organization stores and manages sensitive data itself. On the other hand, if a safe sharing site were to contain data from multiple organizations in order to facilitate a more complex sharing arrangement, this would increase the consequences of a data breach. The exact implementation and related trade-offs to the deployment and use of safe

[72] *See* PRIMAVERA DE FILIPPI & AARON WRIGHT, BLOCKCHAIN AND THE LAW: THE RULE OF CODE 2–3 (2018) (discussing attributes of blockchain).

[73] For a recent discussion of the tensions between blockchain technology and data protection law models, see BLOCKCHAIN AND THE GDPR, THE EUROPEAN UNION BLOCKCHAIN OBSERVATORY AND FORUM (2018).

sharing sites is something we plan to continue to explore in future work. For a visual representation of safe sharing sites, see Figure 1 below.

FIGURE 1: VISUAL REPRESENTATION OF SAFE SHARING SITES



Organization *A*:
Has data

Safe Sharing Site

- Raw data is not disclosed to Organization *B*
- Organization *B* provides code in order to perform computations on data
- Some information, but not the raw data, gets taken out of the safe sharing site

Organization *B*:
Wants data

Legal Interface

Registry with:

- Organization *A*'s claimed lawful authority for processing
- Jurisdiction of the data
- Information on who is using the data and for what purpose

Information in the registry should be auditable.

Organization *C*:
Wants accountability

III
THREE VARIATIONS ON DATA SHARING

In this Part we outline three different examples of data sharing that illustrate how the safe sharing site model can work across different legal contexts. In order to illustrate the legal flexibility of the safe sharing site model, we show how it works in relation to different legal contexts where privacy and data security are key considerations, but which are regulated differently.[74] For each scenario, we outline the basic data-sharing challenge and how a safe sharing site can provide the infrastructure to help solve key problems and enable other potential responses.

[74] We have argued in previous Sections that data sharing sits at a crossroads between privacy and other data governance concerns. However, many of those emerging governance concerns have not yet generated legal regulation and so in these examples we focus on privacy and data security.

A. Ad Auctions

Targeted online advertising largely uses "ad auctions" to deliver ads.[75] In the real-time bidding process (RTB), when an individual loads a webpage that uses RTB, that website communicates with a supply-side platform or ad exchange which then sends out an RTB request. This request broadcasts personal information to other participants in the system (demand-side partners acting on behalf of advertisers) who will then bid on providing an ad to the individual end-user. This practice has long been criticized for its privacy concerns and there are serious legal questions regarding whether it is compliant with stringent data protection law requirements like under the GDPR.[76] Some organizations, including the *New York Times*, have stopped using behavioral targeting in Europe, switching instead to contextual and geographical targeting, because of GDPR concerns.[77]

The individual privacy concerns associated with ad auctions are outlined in two recent complaints that were launched in Europe about the behavioral advertising industry and its compliance with the GDPR.[78] According to the expert report that forms the basis of the complaint, the information that is broadcast in the RTB bid request can include browsing details, location, device information, unique tracking IDs, IP addresses, and data broker segment ID.[79] The latter

[75] *See generally* Lukasz Olejnik et al., Selling Off Privacy at Auction (2013) (considering the privacy implications of different types of ad auctions); Yong Yuan et al., *A Survey on Real Time Bidding Advertising*, Inst. Electrical & Electronics Engineers (2014) (describing the process of real-time bidding advertising).

[76] *See* Ravi Naik, *Grounds of Complaint to the Data Protection Commissioner* (Sept. 12, 2018), https://brave.com/wp-content/uploads/2018/09/DPC-Complaint-Grounds-12-Sept-2018-RAN2018091217315865.pdf [hereinafter *Naik Complaint*] (complaint submitted to the Irish Data Protection Commission); *see also* Ravi Naik, *Submission to the Information Commissioner: Request for an Assessment Notice/Invitation to Issue Good Practice Guidance Re: "Behavioural Advertising"* (Sept. 12, 2018), https://brave.com/wp-content/uploads/2018/09/ICO-Complaint-.pdf [hereinafter *Naik Companion Complaint*] (a companion complaint to the UK Information Commissioner's Office).

[77] *See* Jessica Davies, *After GDPR, the New York Times Cut Off Ad Exchanges in Europe — and Kept Growing Ad Revenue*, Digiday (Jan. 16, 2019), https://digiday.com/media/new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue. Geographical targeting can be based on a user's general location such as country or city, which raises fewer tracking concerns. Contextual advertising is when the advertising is linked to the context of the website that a user visits. For example, if the website is devoted to animals then the advertising would be tailored to this rather than to profiles of individual users. *See* Jessica Davies, *'It's a Balancing Act': Media Buyers Want Contextual Targeting Features to Evolve Further*, Digiday (July 4, 2019), https://digiday.com/media/its-a-balancing-act-media-buyers-want-contextual-targeting-features-to-evolve-further (explaining the rise of contextual advertising in response to data privacy laws).

[78] *See Naik Complaint*, *supra* note 76; *Naik Companion Complaint*, *supra* note 76.

[79] Johnny Ryan, Report from Dr. Johnny Ryan – Behavioural Advertising and Personal Data 4 (2018), https://brave.com/Behavioural-advertising-and-personal-data.pdf.

can include "income bracket, age and gender, habits, social media influence, ethnicity, sexual orientation, religion, political leaning, etc. (depending on the version of 'RTB' system)."[80]

From the perspective of data protection law frameworks such as the GDPR, the core problem of this practice is the lack of control over the conditions of sharing this personal information. The RTB practice incentivizes sharing data with many partners but:

> [E]stablishes no control over what happens to these personal data once an SSP [supply-side platform] or ad exchange broadcasts a "bid request". Even if bid request traffic is secure, there are no technical measures that prevent the recipient of a bid request from, for example, combining them with other data to create a profile, or from selling the data on.[81]

The report claims that "there is no data protection,"[82] but this really means that there is no *effective* protection. There actually is a lot of law involved, including data protection law and industry guidelines and best practices.[83] Players within the RTB system are responsible for complying with the applicable laws and guidelines. The lack of transparency, audit trails, and guarantees regarding safeguards is what leads to the lack of effective protection.

Running an ad auction within a safe sharing site would have a number of benefits. In our scheme, Organization *A* would be a website with the user data and Organization *B* would be a demand-side partner who will want to bid on delivering the ad. *B*'s use of the data from *A* would not involve getting access to the raw data and would instead be undertaken within a controlled environment with the forms of transparency and auditability we have outlined, thereby addressing the "lack of control" arguments. *B* would submit a bidding algorithm that would be run on *A*'s data in the safe sharing site. *B* would only learn whether it won the bid or not, and whether its ad was displayed or not, but would not learn who the ad was displayed to (unless the user who sees the ad clicks on it and in that way directly comes into contact with *B*).

[80] *Id.*

[81] *Id.* at 5.

[82] *Id.*

[83] *See, e.g.*, *The Summary GDPR Compliance Guide for Website*, UniConsent (Apr. 16, 2019), https://www.uniconsent.com/blog/gdpr-compliance-guide-for-website (describing a framework to "help all parties in the digital advertising chain ensure that they comply with the EU's General Data Protection Regulation and ePrivacy Directive"); *see also Authorized Buyers Program Guidelines*, Google (Aug. 22, 2018), https://www.google.com/doubleclick/adxbuyer/guidelines.html (describing the guidelines for Google's Authorized Buyers program, a "service for accessing multiple sources of online display advertising inventory").

The GDPR complaint regarding RTB also raises concerns regarding informed consent and the processing of sensitive information. The requirement for informed consent is a question for each jurisdiction and not something that safe sharing sites address. However, the transparency requirements associated with the legal interface of safe sharing sites can help provide data subjects and regulators with information about the data, including who uses it and how it is used. If this information were available, then it would become easier to establish which parties are engaged in which forms of profiling, and even to use automated tools to track this information at scales not available to individual users.[84]

The concerns regarding sensitive information arise out of obligations under the GDPR regarding "special categories of personal data," which are data that are especially sensitive and linked to fundamental rights and freedoms.[85] In many ways, these concerns are variations on the other concerns regarding safeguards against unauthorized processing and informed consent. However, processing sensitive information creates additional concerns about profiling—inferring sensitive attributes from other data. This is more difficult to address, but safe sharing sites can offer helpful infrastructure to aid in a solution. For example, Organization *A* (the website) could share the data through the site but first require that the safe sharing site provide auditable, technical assurances that: 1) explicit indicators are removed, and 2) processing that would allow Organization *B* to infer sensitive indicators is absent.[86] These assurances would help *A* meet data protection obligations.

Targeted online advertising also raises many issues that go beyond individual privacy. These include the model of "surveillance capitalism" that targeted advertising participates in, the power dynamics involved, the private/public nexus of state surveillance made possible by the data collection, and the broader social effects of commercial trade in our "attention."[87] We are not endorsing the practice

[84] *See* Austin et al., *supra* note 69.

[85] GDPR, *supra* note 20, at art. 9 (including race, ethnicity, political views, religious beliefs, sexual orientation, and genetic information in the special data category).

[86] The adequacy of this depends upon the state of technical research regarding these issues.

[87] *See* Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power 8–12 (2019) (describing the concept of surveillance capitalism); *see also* Bruce Schneier, Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World 78–90 (2015) (describing the "public-private surveillance partnership" between governments and corporations); Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads 5–7 (2016) (describing the commercialization of "attention" and its effects on society).

of targeted online advertising but rather trying to illustrate how—if the practice exists and is accepted—it would be better to implement the practice through a safe sharing site.

B. *Lawful Access*

There are many scenarios where law enforcement wants access to data held by other organizations, such as telecommunications providers or transit authorities. One form this takes is law enforcement seeking information regarding a known target. Many jurisdictions, like Canada and the United States, have well-developed warrant and production order requirements for addressing these requests.[88] Different legal jurisdictions protect these data with different constitutional and statutory standards for access depending on how that jurisdiction understands the privacy interest in the data. For example, the United States largely follows the third-party rule, where data shared with a third party no longer has a reasonable expectation of privacy,[89] whereas Canada has never embraced this rule and has gone so far as to find a reasonable expectation of privacy in basic subscriber information.[90]

Although questions regarding the nature of the privacy interest in different types of data are important, new sharing problems arise in the context of forms of bulk surveillance. These techniques rely upon getting access to large datasets of untargeted individuals so that the datasets can be analyzed in various ways to determine whom to target.[91] In the coverage resulting from the Snowden revelations, these techniques were sometimes discussed as ways of collecting the haystack in order to find the needle.[92] But collecting a haystack of innocent persons' data raises similar privacy issues to the sharing sce-

[88] *See, e.g.*, BAKER MCKENZIE, 2017 SURVEILLANCE LAW COMPARISON GUIDE (2017), https://tmt.bakermckenzie.com/-/media/minisites/tmt/files/2017_surveillance_law.pdf (surveying global surveillance laws and practices); *Global Surveillance Law Comparison*, BAKER MCKENZIE, https://globaltmt.bakermckenzie.com/surveillance-law-comparison-guide (last visited Aug. 28, 2019) (same).

[89] *See* Smith v. Maryland, 442 U.S. 735, 743–44 (1979). *But cf.* Carpenter v. United States, 138 S. Ct. 2206 (2018) (declining to apply the third-party doctrine to the government's use of location information collected from a cell phone database).

[90] *See* R. v. Spencer, [2014] 2 S.C.R. 212, 215 (Can.).

[91] *See* Daragh Murray & Pete Fussey, *Bulk Surveillance in the Digital Age: Rethinking the Human Rights Law Approach to Bulk Monitoring of Communications Data*, 52 ISR. L. REV. 31, 32 (2019) (assessing the use of "bulk communications data surveillance through the lens of human rights law").

[92] *See* Ellen Nakashima & Joby Warrick, *For NSA Chief, Terrorist Threat Drives Passion to 'Collect It All*,' WASH. POST (July 14, 2013), https://www.washingtonpost.com/world/national-security/for-nsa-chief-terrorist-threat-drives-passion-to-collect-it-all/2013/07/14/3d26ef80-ea49-11e2-a301-ea5a8116d211_story.html (quoting an intelligence official as describing the NSA director's approach as "let's collect the whole haystack").

narios already discussed—giving law enforcement access to a haystack reveals the personal information of large numbers of people who are not even under suspicion. This leads to further questions about how those data are safeguarded against abuse.

Cell-tower dumps are an example of bulk surveillance that is perhaps less controversial than national security practices. When cell phones engage in a communication such as making a call or sending a text or email, they connect to their network through the nearest cell tower and a record of this connection is made.[93] These records can provide approximate location information about cell users. Police might want bulk access to cell records in situations such as when they know that a series of crimes has taken place in different locations and they believe that the same person was involved. Cell records can disclose whether any individuals were within close proximity to multiple crime scenes at the relevant time. In the case of an investigation into a single incident, police might want to cross-reference the names of persons accessing a cell tower proximate to a crime scene against the names of owners of vehicle types seen leaving the crime scene.[94]

Cell-tower dumps therefore typically involve either using multiple sets of cell records in order to see whether those sets intersect and show individuals in physical and temporal proximity to more than one crime scene or using a set of cell records that is cross-referenced against other sets of information to determine where those sets intersect. This might reveal an individual in physical and temporal proximity to the crime scene who is also the registered owner of the type of vehicle seen leaving the crime scene.

There are a number of legal questions associated with these dumps, including what kind of authority is required to undertake them. Different jurisdictions can resolve the question of authority in different ways but, regardless, the set intersection analysis should be required to be undertaken in the most privacy-preserving manner possible. In Canada, law enforcement uses a production order.[95] The

[93] For a more detailed explanation of this process, see TAMIR ISRAEL & CHRISTOPHER PARSONS, GONE OPAQUE?: AN ANALYSIS OF HYPOTHETICAL IMSI CATCHER OVERUSE IN CANADA 6–7 (2016), https://cippic.ca/uploads/20160818-Report-Gone_Opaque.pdf.

[94] *See* R. v. Rogers Commc'ns, [2016] ONSC 70, para. 13 (Can. Ont.) (describing typical scenarios).

[95] The production order in *R. v. Rogers Communications* was issued on a "reasonable grounds to believe" standard, which is the same as the reasonable and probable grounds standard in the United States. *Id.* para. 65. However, new types of production orders are now available in Canada for "tracking" and "transmission" data that can be issued on the grounds of "reasonable suspicion." The constitutionality of these new production orders has not been tested in the courts. There is evidence that police rely on these new production orders when using IMSI catchers, which is similar to the cell-tower dump scenario. *See* ISRAEL & PARSONS, *supra* note 93, at 69–70.

Canadian case law suggests that the privacy questions associated with this are largely seen in terms of ensuring data minimization and employing incremental approaches (learn one thing first, then see if you need more information), while safeguards for the data after use are not constitutionally mandated.[96] The courts are in a very poor position to judge questions like data minimization when authorizing a production order. Suggesting, as at least one Canadian case does, that telecommunications companies will undertake to challenge overly broad production orders supports a kind of private sector constitutionalism that also raises many questions.[97]

One prominent proposal for implementing a more privacy-protective protocol for cell-tower dumps is to make use of advances in the ability to do computations on encrypted data.[98] Multiple datasets involving untargeted users can be encrypted and protocols put in place so that the only users who are revealed are those at the intersection of those sets. This protects the private information of untargeted users. Similar methods can also be used for contact chaining, whereby authorities use communication graphs to analyze social connections and identify individuals in groups of interest like criminal organizations.[99]

The proponents of this protocol develop its specifications alongside an argument that law enforcement must follow "open" processes that are public and open for debate.[100] The idea behind their protocol is that the private data of untargeted users can only be processed in encrypted form and then decrypted with specific warrants. The protocol includes involving multiple agencies as a way of ensuring a division of trust and enforcing the limited scope of the envisioned warrants.[101] They also argue that their openness principle requires: 1) a division of trust by having multiple agencies hold encryption keys so that they must cooperate to decrypt the data; 2) that warrants to

[96] *See Rogers Commc'ns*, [2016] ONSC 70, paras. 56, 60 (Can. Ont.) (relying on principle of "minimal intrusion" and leaving to the legislature to address "post-seizure safeguards").

[97] *See id.* para. 2 ("[Communications companies] apply for a court ruling that will make plain that production orders must be tailored to respect the privacy interests of subscribers and conform with constitutional requirements.").

[98] *See, e.g.*, Aaron Segal et al., *Catching Bandits and* Only *Bandits: Privacy-Preserving Intersection Warrants for Lawful Surveillance*, 4 USENIX WORKSHOP ON FREE OPEN COMM. ON INTERNET (2014), https://www.usenix.org/system/files/conference/foci14/foci14-segal.pdf (describing such a protocol).

[99] *See, e.g.*, AARON SEGAL ET AL., OPEN, PRIVACY-PRESERVING PROTOCOLS FOR LAWFUL SURVEILLANCE (2016), https://arxiv.org/pdf/1607.03659.pdf (proposing a "lawful contact chaining protocol" based on the intersections of encrypted data sets).

[100] *Id.* at 2.

[101] Segal et al., *supra* note 98, at 3.

decrypt must be of limited scope; 3) notifying users that their data have been captured; and 4) transparency reporting about the use of mass surveillance.[102] Their technical protocol, therefore, is bound up with normative considerations regarding lawful access.

Implementing protocols like this through a safe sharing site can offer several advantages in terms of oversight and accountability because it offers a way of separating different forms of oversight and accountability from the privacy-protective benefits of the encryption protocols. The basic idea of a safe sharing site is that it can interface with a variety of legal regimes, including those in different legal jurisdictions, which might have different views on the types of authorization required for lawful access and the demands of what Aaron Segal, Bryan Ford, and Joan Feigenbaum call the openness principle.[103] If the encryption protocol for bulk surveillance techniques is tied too closely to the specifics of how particular authorities operate, as well as their technical capacity to participate, then it will be complex to implement both domestically and globally. The best way to implement this kind of privacy-preserving technology is to maintain as much of the current legal infrastructure as possible and offload the technical implementation to the safe sharing site.

C. Litigation

There are a number of litigation scenarios that involve the sharing of data. More scenarios are likely to emerge in the future given the increasing adoption of advanced data analytic techniques to assist, or sometimes automate, areas of human decisionmaking. One example is litigation where one party's claims involve the analysis of a large pool of data and the other party argues that a full response to this claim requires that they be given access to that data for their own independent analysis.

This was at issue in recent Canadian tobacco litigation involving tobacco giant Philip Morris.[104] The province of British Columbia, like many other Canadian provinces, has enacted legislation to enable the province to sue tobacco companies for the health costs associated with tobacco use and incurred by the publicly-funded health care system.[105]

[102] *See id.*

[103] *See id.* at 2.

[104] *See* British Columbia v. Philip Morris Int'l, Inc., [2018] 2 S.C.R. 595, 596 (Can.).

[105] *See* Tobacco Damages and Health Care Costs Recovery Act, S.B.C. 2000, c 30, § 2.1 ("The government has a direct and distinct action against a manufacturer to recover the cost of health care benefits caused or contributed to by a tobacco related wrong."); *id.* § 2.5(b) ("[T]he health care records and documents of particular individual insured persons or the documents relating to the provision of health care benefits for particular individual insured persons are not compellable . . . .").

The legislation barred tobacco companies from having access to individual health records, but Philip Morris argued that they should have access to an "anonymized" version of several of the province's health databases that the province was using to determine causation and damages. Philip Morris lost in the Canadian Supreme Court on statutory interpretation grounds, not because of the broader questions of access to data for the purposes of a fair trial.[106]

The province had offered to make the health care databases available to Philip Morris through a Statistics Canada Research Data Centre. Although some other tobacco companies had agreed to that arrangement, Philip Morris refused the offer because it did not like the conditions imposed on access to the data and argued that the offer involved the waiver of litigation privilege.[107] Philip Morris's complaints included that the analysis and control of the data remained within the control and discretion of Statistics Canada, that Statistics Canada could vet what information left the Centre, and that Statistics Canada required an audit trail that could be obtained by the plaintiff (the BC government).[108]

The tobacco litigation case is a fairly straightforward case of a party wanting access to data for statistical analysis because it is relevant to a fair trial. However, it is also a potential precedent for future litigation involving algorithmic decisionmaking. As others have noted, there are often due process claims that can be made regarding the use of such tools to assist, or automate, human decisionmaking.[109] Often, these claims to transparency and accountability focus on access to the source code for the tool. However, access to source code is likely to be unhelpful in the context of machine learning applications, which can

[106] However, there are still potential future contexts even within the tobacco litigation in which such databases will be compellable. The Supreme Court pointed out that, under the legislation, if the databases are "relied on by an expert witness," then they become compellable, and that the litigants could get access to an anonymized "statistically meaningful sample" upon applying successfully under the statute's requirements. *See Philip Morris Int'l*, [2018] 2 S.C.R. at 614, para. 36 (quoting Tobacco Damages and Health Care Costs Recovery Act § 2.5(b), (d)). The court below had accepted the characterization of the database as anonymous and held that this meant that its compellability did not involve any threat to privacy. *Id.* at 596.

[107] *Id.* at 603, para. 10.

[108] *See* Respondent's Factum, at paras. 14–15, British Columbia v. Philip Morris Int'l, Inc., [2018] 2 S.C.R. 595 (Can.) (No. 37524), https://www.scc-csc.ca/case-dossier/info/af-ma-eng.aspx?cas=37524.

[109] *See, e.g.*, AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 8 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf (discussing due process challenges to the use of algorithmic systems to award government benefits).

remain opaque even to their developers.[110] Access to training data is more helpful for understanding how the model was created and what its potential biases are, although there might also be other ways to make machine learning applications "explainable."[111]

Using safe sharing sites in these litigation contexts would have many benefits. The secure and controlled environment would reduce the risks of unauthorized access or use of the data. But these sites could also be independent from the various parties involved in the litigation, so the problems associated with the proposed use of Statistics Canada Research Data Centres in the tobacco litigation might not arise.[112] The transparency and assurances associated with safe sharing sites can form the foundation of an interface with the court discovery process, permitting court oversight of such data sharing within a litigation context. This might become particularly useful as a piece of infrastructure as legal systems move to develop protocols regarding how to litigate algorithmic fairness issues, as it would provide a secure means for access to the training data of AI systems for those charged with forms of oversight.

IV
Data Governance and Safe Sharing Sites

A. *Safe Sharing Sites and Data Trusts*

The previous Part outlined in detail how safe sharing sites can work in three different legal contexts. Those legal contexts were predominantly concerned with privacy questions, but the applicable legal privacy models differed. However, as previously discussed, in addition to individual privacy, there are many other diverse issues associated with data access and use and a need to develop new governance models to manage those issues. In this Part, we discuss how safe sharing sites can also work in relation to one such model—data trusts. We choose data trusts because they are an interesting emerging model that shares many features with our safe sharing model. We will outline what data trusts are, how safe sharing sites differ, and how the two models could nonetheless fruitfully work together in some contexts.

[110] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. Pa. L. Rev. 633, 638 (2017) (discussing why source code analysis is unhelpful). For an introduction to machine learning, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653 (2017).

[111] *See, e.g.*, Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harv. J.L. & Tech. 841 (2018).

[112] *See infra* Part IV.

The term "data trust" is used in different ways to mean very different things, drawing upon the multiple meanings of "trust," which can range from reliability to a legal property arrangement.[113] For example, in a recent UK report, one recommendation is the "[d]evelopment of data trusts, to improve trust and ease around sharing data."[114] The report uses the term to refer to "proven and trusted frameworks and agreements . . . [that] ensure exchanges are secure and mutually beneficial."[115] Even more recently, the Sidewalk Toronto project announced that it will use a "Civic Data Trust" to manage "urban data" collected in its smart-city proposal based in Toronto, Canada. They appear to mean something more analogous to a legal trust, where a civic organization will manage access to urban data for the benefit of the community.[116]

The Open Data Institute recently documented five different popular usages of the term "data trust": "a repeatable framework of terms and mechanisms"; "a mutual organisation"; "a legal structure"; "a store of data"; and "public oversight of data access."[117] These are all different usages that emphasize, and combine in different ways, what we can think of as three different functions associated with data sharing: decisionmaking regarding access and use; legal compliance and oversight; and the technical mechanism(s) for sharing. Data trusts that are repeatable frameworks or data stores emphasize the technical mechanisms for enabling sharing. The UK report's use of "repeatable framework[s]" is also partly a response to the legal complexities involved in data sharing, emphasizing the need for things like model contracts that can help reduce these complexities.[118] Mutual organizations, legal structures, and public oversight place greater emphasis on models of decisionmaking and legal compliance.

A safe sharing site is more like the data trusts that are a "repeatable framework" or a data store in that it emphasizes the technical mechanisms for data sharing. However, its legal interface is meant to create the necessary infrastructure to work together with different

[113] *See Trust*, ENGLISH OXFORD LIVING DICTIONARIES, https://en.oxforddictionaries.com/definition/trust (last visited May 17, 2019).

[114] HALL & PESENTI, *supra* note 8, at 2.

[115] *Id.* at 4.

[116] *See* Alyssa Harvey Dawson, *An Update on Data Governance for Sidewalk Toronto*, MEDIUM (Oct. 15, 2018), https://medium.com/sidewalk-talk/an-update-on-data-governance-for-sidewalk-toronto-d810245f10f7 (describing the data governance proposal for the Sidewalk Toronto project).

[117] Jack Hardinges, *What Is a Data Trust?*, OPEN DATA INST. (July 10, 2018), https://theodi.org/article/what-is-a-data-trust.

[118] HALL & PESENTI, *supra* note 8, at 46 ("These trusts are not a legal entity or institution, but rather a set of relationships underpinned by a repeatable framework, compliant with parties' obligations, to share data in a fair, safe and equitable way.").

models of decisionmaking or legal compliance and oversight. For example, a safe sharing site does not eliminate the need for data-sharing agreements between the organizations using the site, but it could significantly reduce agreement complexity by taking responsibility for the technical aspects of the data sharing and eliminating the need for the sharing organizations to specify these details. A safe sharing site differs from a data store in that it is not necessarily a repository of data but is instead a mechanism for sharing. In addition, because data stores place the emphasis on particular data sets and their governance, they are focused on specific legal compliance issues. For example, the Silicon Valley Regional Data Trust (SVRDT) integrates data about children from a variety of public agencies in order to "provide a comprehensive understanding of factors contributing to student failure and success."[119] Because of this focus, the SVRDT seeks to integrate specific policy considerations and legal obligations into its technical architecture.[120] In contrast, safe sharing sites are a modular and scalable solution to the issue of data sharing generally, with an interface meant to enable legal compliance across multiple legal contexts.

Because of its modularity, a safe sharing site is not a standalone solution to issues of decisionmaking or legal compliance and oversight. What it does is provide an infrastructure that allows it to work together with other models that accomplish these functions more directly. Some of these models could be data trusts understood as mutual organizations, legal structures, or mechanisms for public oversight. We will briefly outline how this might work in relation to a data trust understood as a legal structure, which is how we understand the "Civic Data Trust" model recently proposed by Sidewalk Toronto.[121]

Legal trusts are a means of holding property for the benefit of some persons or for an object permitted by law. A data trust that is modeled on a legal trust is a potentially flexible instrument for creating an independent body, subject to enforceable obligations, that

[119] Silicon Valley Regional Data Trust, http://www.svrdt.org (last visited May 17, 2019).

[120] *See The Silicon Valley Regional Data Trust: Putting the Last Word First*, Stewards Change Inst., https://stewardsofchange.org/svrdt-putting-last-word-first (last visited May 17, 2019).

[121] As part of the public consultations regarding the Sidewalk Toronto project, Waterfront Toronto and the MaRS Solutions Lab have been drafting a primer on "Civic Digital Trusts." The primer uses the legal model of a trust as its starting point. *See What Is a Trust?*, MaRS, https://marsdd.gitbook.io/datatrust/trusts/what-is-a-trust (last visited May 17, 2019). Note that the language of "Digital Trust" rather than "Data Trust" reflects some feedback that the trust should also govern the digital infrastructure of the project. Since this public discussion is still developing, this Article continues to use the term "Civic Data Trust."

could manage data in the public interest where it falls within regulatory gaps. For example, Canada has relatively strong laws protecting individual privacy, but it lacks a regulatory framework for dealing with governance challenges such as how to deal with data about groups of people or particular neighborhoods, and broader questions regarding data justice and fairness—including the growing concerns around algorithmic fairness and data monopolies. Instead of waiting for the slow process of law reform to create such a regulatory framework, the trust model offers a way of managing these emerging issues through a private law mechanism. Like the idea of a safe sharing site, a legal trust is a flexible piece of legal infrastructure that is adaptable to many different uses.

However, a data trust understood as a legal trust still requires a means of sharing data and thus still involves issues like re-identification risks and the general need for transparency and auditability. Sidewalk Toronto's data trust proposal appears to rely upon the "open data" model that we have criticized in this Article—although it is still under development. But a data trust could instead integrate well with a safe sharing site. Within our schema, the data trust would become "Organization *A*" and manage access and use decisions regarding the data it controls. The sharing would occur through the safe sharing site, which would operate independently of the data trust. This independence is a benefit if municipalities are considering how to make investments to enable smart city projects. A Civic Data Trust can use a safe sharing site, but so can a municipal government that wants to share data (that is not held in the trust) with a private sector organization for a particular purpose, or two private sector organizations who would like to share data that is not held in the trust. As such, a safe sharing site can be an important component of legal-digital infrastructure for the data economy that can both work together with a Civic Data Trust but also has independent value.

The legal trust model works well in conjunction with safe sharing sites, since the trust provides a model for decisionmaking in the public interest and the safe sharing site provides a reliable mechanism for implementing that decisionmaking. However, there are a few potential legal hurdles to the legal trust model for data trusts. One problem concerns whether the data to be held in trust can be the subject of property rights.[122] Another problem might be enforceability. The data trust model is closer to a non-charitable purpose trust than to a reg-

[122] Data can only be owned in limited circumstances. *See generally* TERESA SCASSA, CTR. FOR INT'L GOVERNANCE INNOVATION, CIGI PAPERS NO. 187, DATA OWNERSHIP (2018), https://www.cigionline.org/publications/data-ownership (describing the current legal terrain that governs data ownership).

ular trust and such purpose trusts have historically faced legal challenges based on the lack of an enforceability mechanism.[123] It might be that overcoming such legal impediments will require enabling legislation. A discussion of this is beyond the scope of this Article.

B. Governance Challenges of Safe Sharing Sites

Safe sharing sites themselves face a number of governance challenges. Technological solutions alone cannot address all problems of potential misuse of data by the operator of the site. There need to be laws imposing liability for data misuse and oversight mechanisms to enforce them. Moreover, if safe sharing sites are to develop as a kind of legal-technical infrastructure that facilitates data sharing across diverse use contexts, then there is a need to develop technical standards. These standards are required in relation to the privacy-protective modes of data sharing envisioned, as well as for the legal interface and its registry that is so essential in enabling various forms of regulation and governance.

One way to address the questions of data misuse and oversight is to implement safe sharing sites within existing regulatory frameworks. For example, a recent proposal from the Toronto Region Board of Trade in relation to the Sidewalk Toronto project is to create a "Data Hub" to be managed by the Toronto Public Library and overseen by the Information and Privacy Commissioner of Ontario.[124] Something similar could be proposed for safe sharing sites generally. Safe sharing sites can use strong preexisting data protection laws and established networks of data regulators to address concerns regarding data misuse.

The problem with this approach, however, is that it relies heavily on existing data protection law. This places concerns about data misuse within the paradigm of individual privacy that underpins that

[123] *See* Richard C. Ausness, *Non-Charitable Purpose Trusts: Past, Present, and Future*, 51 REAL PROP., TR. & EST. L.J. 321, 328 (2016) (noting that for much of the past two hundred years, courts often declined to uphold non-charitable purpose trusts "because there was no human beneficiary to enforce them"). Ontario allows non-charitable purpose trusts in limited circumstances but only through the operation of a saving provision that converts them into powers that are not enforceable. *See* Perpetuities Act, R.S.O. 1990, c. P.9 § 16 (Can.). This is in contrast to some other jurisdictions. For example, in the United States, the Uniform Trust Code endorses purpose trusts. *See* UNIF. TRUST CODE § 409 cmt. (UNIF. LAW COMM'N 2010). Other examples include Guernsey, which explicitly permits the creation of non-charitable purpose trusts and has statutory provisions for enforcement. *See* Trusts (Guernsey) Law § 12(1) (2007).

[124] The form that this "Data Hub" will take has not been specified, with suggestions that it could be a "Data Trust" or a "Data Repository." *See* TORONTO REGION BD. OF TRADE, BIBLIOTECH: BEYOND QUAYSIDE: A CITY-BUILDING PROPOSAL FOR THE TORONTO PUBLIC LIBRARY TO ESTABLISH A CIVIC DATA HUB 1–2 (2019).

form of regulation. Even from a privacy perspective, this is problematic because, as we outlined previously, privacy concerns can arise in contexts that are not regulated by data protection law. If safe sharing sites are also supposed to help address a broader set of emerging data governance concerns, whether directly or indirectly, then it is even more unclear that placing their oversight within a data protection law framework makes sense.

Ideally, safe sharing sites would be regulated on their own and in a manner that allows for the reduction of regulatory complexity. For example, suppose organizations wish to undertake data sharing that is governed by three different regulatory regimes that address privacy concerns with three different regulators. If each of these regimes allowed organizations to discharge a subset of their obligations (those pertaining to privacy and data safeguards in relation to sharing) through the use of a certified safe sharing site, then the legal complexities for organizations involved in data sharing would be reduced. If the obligations and standards of the safe sharing site were regulated independently of these three regimes and three regulators, then the legal complexities for the safe sharing sites would be reduced. This would be a governance model that matches the modular nature of the safe sharing site. The idea is that there are core problems associated with data sharing that can be solved independently of the contexts in which this sharing occurs even though this solution must still interface with those different contexts.

These challenges regarding oversight have been raised in other, similar, discussions. For example, the UK report proposes the creation of a Data Trusts Support Organisation (DTSO) to help oversee the development of Data Trusts.[125] Something comparable to this might be needed in order to facilitate the creation of a global network of safe sharing sites. Other interesting models include Gillian Hadfield's proposal to create more market incentives for regulation, whereby governments set general objectives and allow for private non-profit or for-profit regulators to gain certification and build the specific regulatory processes.[126] A full discussion of the merits of these alternatives, and others, is beyond the scope of the present Article.

The second, related set of challenges concerns the development of standards. Some of these standards will be technical in nature. Protecting against access to the raw data is one of the core functionalities of a safe sharing site and is an important aspect of privacy protection.

[125] HALL & PESENTI, *supra* note 8, at 47.

[126] *See* GILLIAN HADFIELD, RULES FOR A FLAT WORLD: WHY HUMANS INVENTED LAW AND HOW TO REINVENT IT FOR A COMPLEX GLOBAL ECONOMY 4 (2016) (providing an overview of her proposal).

Safe sharing sites would keep the raw data—which includes PII—from ever leaving the site. However, it cannot prevent all information from leaving the site or this would defeat its purpose. Organization *B* wants access to Organization *A*'s data in order to perform some kind of computation on it and will need to extract some kind of output from this computation. This can lead to new kinds of privacy challenges. One example comes from machine learning research: Researchers have shown that machine learning models can be "inverted" to extract information about the individual samples in the training set.[127] In other words, even if someone does not have access to the data used to train a machine learning model, they may still be able to learn something about the data. To defend against this, researchers have turned to privacy-preserving tools that will transform the dataset so that private information can be less easily extracted, regardless of the machine learning methodology applied.[128] This approach is attractive because the privacy-preserving mechanism is independent of the machine learning algorithm, enabling strong guarantees regardless of the machine learning technique applied. However, the drawback is that this conservative approach decreases the accuracy of the machine learning classifier that can be trained. This has led to further attempts to integrate differential privacy directly into the machine learning training algorithm.[129]

Other standards will be needed to create the legal interface of safe sharing sites. The registry mechanism that we envision will work best if the data from these registries is standardized across different sites. As previously outlined, this is not necessarily "public" data, but it is data that can be made available to regulators and oversight bodies when appropriate. An element of enabling effective regulation in relation to an increasingly complex data ecosystem is to develop automated tools to help regulators access this registry information and use it to map data-sharing activities and flag potential problems for further investigation. Safe sharing sites are also, ideally, meant to operate globally. This means that jurisdictional information will also be required. This is important information for the registry because it will

[127] *See* Matt Fredrikson et al., *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, 22 ACM CONF. ON COMPUTER & COMM. SECURITY 1322, 1332 (2015).

[128] *See, e.g.*, Martín Abadi et al., *Deep Learning with Differential Privacy*, 23 ACM CONF. ON COMPUTER & COMM. SECURITY 308, 308 (2016).

[129] *See, e.g.*, Nicolas Papernot et al., *Scalable Private Learning with PATE*, 2018 INT'L CONF. ON LEARNING REPRESENTATIONS (describing research).

help different regulators or oversight agencies understand whether the activities of the safe sharing site fall under their jurisdiction.[130]

Finally, there are a set of concerns associated with ensuring that safe sharing sites can work in the public interest and do not inadvertently form a costly barrier to data sharing that only commercial entities can access. For example, in the freedom of information context, Margaret Kwoka has documented how the quality and character of commercial requests can negatively affect the ability of FOIA regimes to achieve their ultimate goal of governmental accountability.[131] Data sharing, on our account, serves many ends, but it should not be the case that mechanisms to enable safe sharing become a source of significant transaction costs that hinder the ability to use public data.

Conclusion

One of the lessons of the Sidewalk Toronto project is that data sharing is an essential activity within the data economy and also sits at the crossroads between individual privacy concerns and broader emerging issues in data governance, including ensuring public benefit and control, resistance to data monopolies, social surveillance and harms, and algorithmic fairness and accountability. As we have argued, data sharing as a practice can be the cause of these concerns at the same time that it can be a solution to these concerns.

Our safe sharing site approach to data sharing focuses on resolving key risks associated with data sharing, including protecting the privacy and security of data subjects, but does so in a manner that is independent of the various legal contexts of regulation and governance. We propose that safe sharing sites connect with these different contexts through a legal interface consisting of a registry that provides transparency about key information that supports different forms of regulation. Safe sharing sites also offer assurance and auditability regarding the data sharing, further supporting a range of regulatory interventions. It is therefore not an alternative to these interventions but an important tool that can help enable effective regulation.

[130] Some kind of tagging protocol would have to be developed and used consistently in different jurisdictions. This also involves questions regarding how to tag the data. Jurisdiction information could refer to the location of collection, it could refer to the location of the data subject, and it could be based on the citizenship or residency status of the data subject.

[131] *See* Margaret B. Kwoka, *FOIA, Inc.*, 65 Duke L.J. 1361, 1414–26 (2016) (noting that commercial FOIA requests act as a form of unintended corporate subsidy, overwhelm government resources, and transform certain companies into "brokers of public information").

In this way, safe sharing sites facilitate data sharing in a manner that manages the complexities of sharing while reducing the risks and enabling a variety of forms of governance and regulation. As such, the safe sharing site offers a flexible and modular piece of legal-technical infrastructure for the new economy.