



# THE ETHICAL REUSE OF DATA IN A MACHINE LEARNING WORLD

REPORT FROM AN OCTOBER 2017 WORKSHOP HOSTED BY  
GEORGETOWN LAW'S INSTITUTE FOR TECH LAW & POLICY

**Aaron Fluitt, Fellow**  
**Alexandra Reeve Givens, Executive Director**  
**Georgetown Institute for Tech Law & Policy**

---

# TABLE OF CONTENTS

<b>PART ONE: Emerging Themes Surrounding Data Reuse in the Machine Learning Age</b>	2
A. Characterizing Data Reuse in the Machine Learning Context	2
1. Supervised Learning Versus Unsupervised Learning	2
2. Increases in Analytical Capacity & Access to Data Mining Tools	2
3. Machine Learning Can Discover Non-Intuitive Correlations	3
4. New & More Complicated Questions for Old Regimes	4
B. Identifying Tensions Between Fair Information Practices & Data Reuse in the Machine Learning Age	
1. Challenges of Adhering to the Collection Limits Principle	6
2. Challenges of Adhering to the Purpose Specification & Use Limitation Principles	6
3. Challenges of Transparency in General	7
4. The Future of the FIPs	7
<b>PART TWO: The Competing Costs &amp; Benefits of Data Reuse</b>	
A. Balancing Benefits & Risks in Determining Whether to Permit Data Reuse	9
1. Identifying, Classifying, & Quantifying the Benefits & Risks of Data Reuse	9
2. Risks & Harms of Data Reuse	11
B. Identifying & Defining the Political Values, Rights & Principles at Stake in Reuse Scenarios	12
1. Privacy	12
2. Data Protection	12
3. Addressing Bias & Inequality	13
4. Enabling & Promoting High-Quality Scientific Research	14
5. First Amendment Rights & Values	14
6. Serving Communities & People in Need	15
<b>PART THREE: Regulating Data Reuse in a Machine Learning World</b>	
A. Regulating Collection & Consolidation of Data	16
1. Regulating Data Collection	16
2. Regulating Data Consolidation: Data Pools vs. Data Lakes	16
B. Regulating Use & Access to Data	18
1. Characterizing & Classifying Purposes & Uses of Big Data	18
2. Privileged Access vs. Open Data	18
C. Assigning Responsibility for Regulating Data Reuse	20
1. Belmont-Plus & IRB-Light Models	20
2. Adapting Traditional Legal Mechanisms	24
3. Trusted Entities Model	24
<b>PART FOUR: Topics for Future Discussion</b>	
A. Is Differential Privacy “Privacy”?	26
B. Open Data, Privileged Access, & Data Clearinghouses	26
C. Identifying & Weighing Benefits & Risks	27
<b>PARTICIPANTS</b>	32

# INTRODUCTION

Recent advances in machine learning and other forms of artificial intelligence have spurred a great data rush. Because machine learning techniques rely on massive troves of data to train and refine powerful models of prediction and decision-making, these techniques serve up a clash of moral imperatives: the need to protect privacy and the need to solve difficult societal problems.

This report summarizes the presentations and reflections of leading experts from academia, computer science, government agencies, public interest groups and private companies at a workshop exploring this intersection hosted by Georgetown Law.<sup>1</sup> The workshop focused on identifying and confronting the ways in which big data and machine learning appear to be challenging traditional legal, ethical, and attitudinal approaches to limiting the reuse of data.

The workshop focused on the primary question of when, if ever, it may be appropriate to share or reuse data that was initially gathered for a different purpose. The workshop began with an overview of machine learning systems and the Fair Information Practices (FIPs), which for decades have formed the backbone of data privacy regulations around the world. Participants debated whether the FIPs remain a practicable framework for analyzing privacy in the context of big data and machine learning operations on certain data sets. The workshop then focused on distilling points of consensus and contention on the future of data privacy in an era of big data and machine learning.

Ultimately, the statements of participants in the workshop demonstrate that the questions raised by data reuse in a machine learning world are important to broad segments of society, but also difficult to resolve given current legal and institutional frameworks, making this area ripe for further analysis and research.

---

<sup>1</sup> Roundtable, *Workshop on the Ethical Reuse of Data in a Machine Learning World*, INSTITUTE FOR TECHNOLOGY LAW & POLICY, GEORGETOWN UNIVERSITY LAW CENTER (Oct. 27, 2017). The roundtable workshop was an invite-only dialogue with participation under the Chatham House Rule to encourage free and frank discussion of the issues.

# PART I. EMERGING THEMES SURROUNDING DATA REUSE IN THE MACHINE LEARNING CONTEXT

## A. CHARACTERIZING DATA REUSE IN THE MACHINE LEARNING CONTEXT

A threshold question in determining the appropriate framework discussing data reuse in the machine learning context is whether machine learning’s use of data meaningfully differs from more traditional statistical analyses. Is the difference between machine learning and traditional techniques one only of *degree*—in the quantity and availability of data and the capacity for analysis—or of *kind*, in that machine learning processes differ so significantly that we must think of an entirely new approach to data protection and privacy?

### 1. Supervised Learning Versus Unsupervised Learning

Several participants noted that the distinction between supervised and unsupervised machine learning may be important to the characterization question. In both supervised and unsupervised machine learning systems, researchers provide the system with “training data”—data from which the system can uncover underlying correlations between variables, which it then uses to generate an algorithm capable of making predictions about future cases. In supervised learning, researchers must actively label or classify examples of target variables of interest.<sup>2</sup> Unsupervised learning, on the other hand, “do[es] not require any such target variables and instead search[es] for general structures in the dataset, rather than patterns specifically related to some state or outcome.”<sup>3</sup> The system is essentially free to discover whatever correlations it can. Several participants noted that unsupervised learning seems to represent a significant departure from traditional statistical analysis that would represent a difference in kind.

In the context of supervised learning, several participants contended that the difference is a matter of degree; a difference primarily in the quantity and availability of data that can now be analyzed—due in part to the sheer amount of data being captured in individuals’ daily lives, and the ease of accessing such pools of data through powerful search engines, cross-indexed databases, and the accumulation of data online.

### 2. Increases in Analytical Capacity & Access to Data Mining Tools

Another difference is the question of *who* may process large quantities of data. Increasingly, technology is making it easier for a broader group of people to conduct extensive data analysis, both

---

<sup>2</sup> Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CAL L. REV. 671, 678 n.24 (2016).

<sup>3</sup> *Id.*

because more companies and entities have, or can purchase, access to the data, and because computerized tools are more accessible than traditional statistical techniques.<sup>4</sup>

One participant commented that the capacity to perform the sorts of analyses that machine learning makes possible “is now much greater than it used to be, but the idea that you would have a bunch of information, and make generalizations about the group of people who you have information about, and then apply it to a new case, [is] not new in principle, [it’s] called social science.” The participant referred to credit scoring algorithms as one example of “old-fashioned algorithms” that operate similarly to supervised machine learning systems.

### 3. Machine Learning Can Discover Non-Intuitive Correlations

Another participant remarked that the credit scoring example, which was held out as evidence that machine learning represents only a difference in degree, actually highlights a real-world example of how even supervised machine learning could be different in kind. A machine learning system operating on a large set of financial data to come up with a better methodology for credit scoring might “discover as a byproduct . . . a way of identifying bad drivers.” This example raised two points where some participants suggested machine learning differs significantly from traditional research.


First, the ability of machine learning systems to discover unintuitive and surprising patterns means that researchers will not always know what the purpose or anticipated output of a system will be at the outset. This difference renders machine learning techniques particularly discordant with traditional approaches to Fair Information Practices (FIPs), which require that data subjects be given notice of how their data will be used and the purposes to which it will be put.

Second, the example highlights the challenges created by machine learning’s power to identify correlation without necessarily giving insight into causation (in the hypothetical, *why* credit scores would correlate with bad driving). For several participants, this “black box” nature of machine learning algorithms indicates a significant distinction from traditional statistical analyses, potentially creating novel public policy issues regarding oversight and the ability to evaluate the social justice impacts of such systems.

On the other hand, another participant remarked that long before the recent rise in machine learning systems, automobile insurers discovered that credit scores were an effective predictor of

---

<sup>4</sup> For example, analytics company AirDNA scrapes the “information publicly available on the AirBnB website” and uses machine learning technology to generate and sell custom data reports and an “interactive market intelligence tool.” *AirDNA Data Methodology*, <https://www.airdna.co/methodology>. AirBnB itself offers a free and open-source machine learning tool called Aerosolve meant to provide price tips to hosts, but its engineers have suggested other uses as well, such as “teaching the algorithm how to paint in the pointillism style of painting” and predicting household income “based on US census data.” Hector Yee & Bar Ifrach, MEDIUM, *Aerosolve: Machine learning for humans*, <https://medium.com/airbnb-engineering/aerosolve-machine-learning-for-humans-55efcf602665>.



the amount of claims an individual would file and virtually every automobile insurer used credit scores to determine car insurance premiums. The FTC, the participant noted, did a study on insurers' use of credit scores, and while acknowledging the scores were an effective predictor of insurance risk, the FTC could not "figure out what it is that the credit score is keying into."<sup>5</sup> Partly because of this mystery, the participant said, and partly because the credit scoring model had a "differential impact on racial groups," some state regulators prohibited the use of credit scores for insurance purposes.<sup>6</sup> While contending that this sort of opacity in the operations of predictive modeling systems has been an issue for regulators for decades, the participant agreed that big data and machine learning will increase the frequency and accelerate the appearance of these transparency issues in many new contexts.

#### 4. New & More Complicated Questions for Old Regimes

Several participants remarked that machine learning presents a host of complicated new questions that significantly impact applicable privacy regimes.

First, substantial increases in data quantity and ease of access carry increased risks that ostensibly anonymized datasets may be combined and cross-referenced to discover the identity of individual data subjects.<sup>7</sup>


Second, big data and machine learning may exacerbate what one participant referred to as a "tyranny of the minority" situation, whereby the voluntary data sharing practices of a small number of individual data subjects can be studied to produce predictive models which end up affecting large populations of individuals who did not participate in the study. For example, a relatively small number of Facebook's billions of users are entirely comfortable publicizing generous amounts of information about themselves that most people consider private. That fraction of Facebook users could provide researchers with sufficiently massive datasets to enable machine learning systems to make correlations between those typically personal details and other information more traditionally shared on social media. Those correlations could then be used to draw inferences from the general information shared by a majority of users about those sensitive details which most never intended to make public.

---

<sup>5</sup> See Press Release, FTC, FTC Releases Report on Effects of Credit-Based Insurance Scores (July 24, 2007), <https://www.ftc.gov/news-events/press-releases/2007/07/ftc-releases-report-effects-credit-based-insurance-scores>. Report available at [https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta\\_report\\_credit-based\\_insurance\\_scores.pdf](https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf).

<sup>6</sup> The participant noted that other states allowed insurers to use credit scores but only if they provided the source code, highlighting one regulatory approach to handling the transparency issue that often arises in the machine learning context.

<sup>7</sup> This issue has been written about extensively, including by one of the conveners of the workshop, Paul Ohm, in *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).



By way of example, the participant referred to “recent papers showing that people had scraped social networks to get images of transgendered peoples’ faces, or scraped social networks to get the faces of ‘out’ gay people, and then use that as training data to build a model to make inferences about other people.”<sup>8</sup> The participant noted the risk of entities pursuing a divide-and-conquer strategy, where researchers could almost always find a sufficient number of people who are willing to share information that most people would not want to disclose, and that “small group of people would then dictate for the entire population.” However, the participant noted that while it still seems like such a situation could arguably result in widespread harms, it is also at the very core of the purpose of research, to learn new information by researching small samples of the population which can then be applied generally to the entire population.

The participant also pointed out a third point raised by the examples of scraping photos from public social media, regarding the growing gap between people’s intuitions about the privacy of their information and the practical realities of machine learning systems. People sharing their photos on Facebook, the participant noted, did not actively participate in the research process and may not have even realized that their pictures could be useful in research of that nature. The participant remarked that cases like these feel very different in that “the data we can now use to make a reasonably decent prediction or inference about people is increasingly non-intuitive and distant from what people’s intuitions might be.”

## B. IDENTIFYING TENSIONS BETWEEN FAIR INFORMATION PRACTICES & DATA REUSE IN THE MACHINE LEARNING AGE

The FIPs have governed the collection and use of data for decades, in short by requiring transparency in the collection and use of data, usually through provision of notice to potential data subjects, and the solicitation of consent from individual data subjects as preconditions to the collection, use, and reuse of individuals’ data. Participants’ discussion of the FIPs began with a brief overview of three FIPs principles relevant to big data reuse in a machine learning age: collection limits, purpose specification, and use limitation. Participants discussed the challenges big data researchers face when attempting to adhere to the FIPs. They debated whether these tensions suggest that the FIPs themselves fail us in the machine learning age, or whether the tensions are primarily the result of researchers’ failure to adhere to the FIPs.

---

<sup>8</sup> See, e.g., Michael Kosinski & Yilun Wang, *Deep Neural Networks Can Detect Sexual Orientation from Faces*, draft available at <https://osf.io/fk3xr/>; see also, Heather Murphy, *Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine*, N.Y. TIMES, Oct. 9, 2017, <https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html>.

## 1. Challenges of Adhering to the Collection Limits Principle

The collection limits principle counsels that “companies should collect only as much personal data as they need”<sup>9</sup> to accomplish specified purposes, and that personal data ought be obtained “by lawful and fair means, where possible, with the knowledge or consent of the data subject.”<sup>10</sup> One participant noted that in the context of big data reuse, there may often be a lack of subjects’ knowledge and consent with regards to collection of their data, even when obtained lawfully.

Another participant referred to the example of recent research conducted by scraping photos from publicly available social media, where even though individuals might post photos, they may be unaware of the fact that researchers can access and collect their photos.

Participants did not come to a conclusion about whether notice and consent of these sorts of secondary collection practices is possible in a machine learning world, but several participants argued that individualized notice and consent at the moment of secondary collection would be impractical, and one participant suggested that collection and use has for some time been so complex that in many cases giving notice and obtaining consent may not be feasible or even possible.<sup>11</sup>

## 2. Challenges of Adhering to the Purpose Specification & Use Limitation Principles

The purpose specification principle requires that data subjects are notified of the purposes to which their data will be put at the time of collection. In addition to raising issues relating to cases of secondary collection discussed above, participants also pointed out that purpose specification can be problematic in the machine learning context when the ultimate goals or outcomes of big data research are not always clear at the outset.


One participant noted that the purpose specification principle includes a clause which does allow for the reuse of data for other purposes “as are not incompatible with” those purposes initially specified. The participant remarked that earnest efforts to define the boundaries of that clause have proved unsuccessful.

---

<sup>9</sup> Robert Gellman, *Fair Information Practices: A Basic History* at 25 (2017), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2415020](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020).

<sup>10</sup> *Id.* at 7.

<sup>11</sup> “In the education example, data is coming from educational institutions, but it is data about individual students, so the pattern to get that notice down to the individual students is going to be a very complex relationship that will dilute the ability to have something meaningful.”



Relatedly, the use limitation principle prohibits disclosure or use beyond those specified purposes, “except with consent . . . or by authority of law.”<sup>12</sup> Because the workshop’s focus was on data *reuse* for *novel* purposes, both the purpose specification and use limitation principles highlighted the main points of tension between the FIPs and machine learning systems.

### 3. Challenges of Transparency in General

While some workshop participants highlighted areas where machine learning systems pose increased problems for transparency, all who spoke about this point seemed to agree that the fundamental notice and choice framework of the FIPs had in many respects broken down long before the rise of machine learning. However, participants differed as to whether that breakdown was attributable to some inherent shortcoming of the FIPs themselves rather than a failure of industry compliance with, or regulatory enforcement of, the FIPs.

One participant suggested that the workshop recognize industry’s role in creating the problems that industry is now “hankering to fix.” The participant noted that “it’s no surprise that notice and choice is broken when extraordinarily few entities in industry have any incentive to create robust, meaningful, notice and choice regimes.”

One participant suggested that the fact that potential future uses are often unknown and potentially unknowable at the time of collection is neither entirely new nor exclusive to machine learning. Providing “notice” of the potential broad uses of certain data often requires that notice be given with such generality as to be effectively meaningless.


Moreover, a few participants noted that from the outset, notice and choice has been riddled with exceptions that often come to rest on marginalized populations; for instance, participants in social welfare programs are often required to waive any control they might have had over their data in order to be eligible for the benefits of these programs. In addition, information on members of “migrant populations and people who encounter the national security state in various ways” has been included in “databases that are born into the world outside of the notice and choice framework,” because national security has always acted as an override of FIPs where transparency and notice and consent are antithetical to the nature of the national security state.

### 4. The Future of the FIPs

Ultimately, participants did not arrive at a consensus regarding the FIPs’ future prospects for protecting the rights of data subjects.

---

<sup>12</sup> Robert Gellman, *Fair Information Practices*, *supra* n.9 at 7.



At least one participant seemed to suggest that the FIPs might be able to effectively protect data subjects while permitting big data analyses if only the proper incentives were implemented to ensure entities' meaningful compliance with the FIPs.

Several participants who spoke seemed to agree that the FIPs must be revisited. And at least one participant seemed to suggest that the FIPs ought to be discarded in favor of some other framework for determining whether particular reuses of data ought to be permitted. Some participants who expressed skepticism of the FIPs argued that, whereas the value choices embedded in the FIPs are centered upon the determinations of an *individual data subject* (notice and consent), big data and machine learning make such reliance on the determinations of an individual subject impracticable. By necessity, a more functional framework may instead need to focus on balancing broader considerations of public and private benefits and costs. This brings with it its own complexities, as discussed in more detail below.

---

## PART II. THE COMPETING COSTS AND BENEFITS AROUND DATA REUSE

### A. BALANCING BENEFITS & RISKS IN DETERMINING WHETHER TO PERMIT DATA REUSE

If concerns about the reuse of data require us to move from the FIPs' focus on individualized notice and consent, we might turn instead to a balancing of benefits and costs in reusing sets of data. This raises two key questions: how should that balancing decision be made, and who gets to make it? Embedded within the question of "how the balancing decision should be made" are additional questions about who gets access to the data, under what conditions (e.g. subject to what security safeguards), and for what purpose? Embedded within the question of "who gets to make it" is the choice of a governance framework—relying on individual decision-makers, industry best practices, or a stronger regulatory approach?

One threshold problem in the "balancing test" approach is that, because the benefits of data reuse in the machine learning context may be unknowable at the outset of a data study, it may be impossible to engage in a full balancing of benefits and risks to determine whether data reuse ought to be permitted in any given instance.

More generally, participants identified several persistent problems plaguing the regulation of traditional statistical analysis and research which carry over to the machine learning context, including problems related to identifying, classifying, and quantifying not only the benefits but also the risks of data reuse, which in turn raise questions regarding the identification and definition of the values, rights, and principles at stake in data reuse scenarios.

#### 1. Identifying, Classifying, & Quantifying the Benefits & Risks of Data Reuse

The workshop highlighted several challenges involved in identifying precisely the benefits and risks of data reuse, and the lack of public consensus regarding the relative weight of particular benefits and values. To highlight the complexities involved, one participant invoked the case of GPS manufacturer TomTom's sharing of user-generated driving data with government entities. At least one local law enforcement agency used TomTom's data to figure out where to place speed traps. Because the public objected to the way the city used speed traps—as a way to raise money rather than to reduce speed-related accidents—public outcry precipitated a change in TomTom's data sharing practices. The participant explained the nuances of assessing public benefit:

*To use [the] TomTom example, it would be one thing if the municipality used this to address an intersection where there are lots of people killed because of traffic accidents. . . . [T]hen there [are] public non-health benefits. What if the municipality used the data to basically ameliorate traffic that isn't killing people but is just slowing them down? . . . There are different ways to achieve that benefit. One way is to put up traffic cameras and ticket people, another way is to put up speed bumps that just slow people down but don't actually harm*

*people in any financial way. And then finally, [a contrasting use] would be if you sold the data to Starbucks so they could set up Starbucks at highly trafficked corners in a way that really only benefits Starbucks. I think we need to interrogate [the question of benefits and harms] a lot more.*

#### *a. Benefits of Data Reuse*

In addition to the aforementioned difficulties of specifying potential benefits of data reuse in machine learning systems *a priori* due to the often dynamic, speculative, and unintuitive results of such reuse, workshop participants also discussed problems in distinguishing between public and private benefits, the disparities between actual and perceived benefits, and the likely lack of public consensus on the relative valuation of distinct classes of benefits.

#### *b. Public vs. Private Benefits*

A recurring theme centered around whether certain benefits should be thought of as accruing to the general public or to individual entities. Some participants noted that the distinction between public and private benefits is not always clear. For instance, if a hospital or university is reusing data it collected in order to improve outcomes of patients or students, then in one regard it will be the public who benefits through better medical or educational outcomes. However, the institutions themselves also receive a benefit through reputational improvements and increased demand for their services.

Another question around the public-private benefit dichotomy asked whether individual data subjects should be entitled to some benefit, financial or otherwise, in exchange for the reuse of their data for a novel purpose. Some participants noted that this would likely create the same sorts of difficulties that undermined the requirement of individualized consent in the context of FIPs.

#### *c. Distinguishing Actual & Perceived Benefits*

One workshop participant noted that there is often a distinction between actual and perceived benefits, and that the concept of “benefit” may not be the most helpful metric by which to make decisions evaluating when and how to permit data reuse. For instance, while in the context of institutions like hospitals or universities the benefit may be regarded as a public benefit—better outcomes for all patients or students—individual patients and students may misconstrue the metric to ask why their data should be shared when they do not “benefit” themselves. Because the data subjects and the beneficiaries of the data reuse are often not the same individuals, a focus on “benefit” can obfuscate the broader balancing. A more helpful framework may be to speak in terms of “social good,” a term often used in the environmental context where the same problem of societal vs. individual benefit also emerges with frequency.

## 2. Risks & Harms of Data Reuse

Just as the benefits of data reuse may be challenging to quantify, participants noted that quantifying harms can prove similarly challenging. One participant wondered whether the mere reuse of data without individualized consent ought to constitute a harm in itself. This approach embraces the FIPs' central premise that individualized notice and consent remain important as mechanisms for protecting individual rights. The question then becomes whether such a harm can be weighed against the potential benefits of reusing a person's data.

In further discussions of measuring harm, some participants suggested that harms ought to be defined as the taking of some specific action with regard to an individual data subject; for example, the denial of a loan. Then, bearing the example of denials of loan applications in mind, the question of public vs. private harms was discussed. The potential "tyranny of the minority" situation, where individuals who were not subjects of the data sets are nonetheless negatively affected by knowledge accrued from the use of consenting subjects' data, raised the question of whether harms to individuals whose data was not relied upon ought to be taken into consideration in any cost-benefit analysis.

Some participants questioned whether privacy losses and invasions would continue to present serious risks of harm in the machine learning context. At least one participant noted that novel approaches to protecting privacy, such as frameworks or mechanisms that satisfy the concept of differential privacy,<sup>13</sup> might substantially mitigate or even effectively eliminate the risks of re-identifying individual data subjects which have heretofore been exacerbated by big data. Another participant remarked that the notion of privacy itself has not been amenable to a clear definition, such that the question of how to define privacy and privacy loss has not settled on any general consensus. As one participant pointed out, there are likely to be stark differences between the expectations and intuitions of the general public, and the legal, formal, and practical realities of what privacy actually entails.

---

<sup>13</sup> Cynthia Dwork, *Differential Privacy*, 4052 AUTOMATA, LANGUAGES & PROGRAMMING 1, 1–12 (2006), (Michele Bugliesi et al. eds., 2006), [https://link.springer.com/chapter/10.1007/11787006\\_1](https://link.springer.com/chapter/10.1007/11787006_1); see also Cynthia Dwork, Frank McSherry, Kobbi Nissim, & Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, 3876 PROCEEDINGS OF THE THIRD CONFERENCE ON THEORY OF CRYPTOGRAPHY 265, 265–84 (Shai Halevi & Tal Rabin, eds., 2006), [https://link.springer.com/chapter/10.1007%2F11681878\\_14](https://link.springer.com/chapter/10.1007%2F11681878_14).

## B. IDENTIFYING & DEFINING THE POLITICAL VALUES, RIGHTS, & PRINCIPLES AT STAKE IN DATA REUSE SCENARIOS

One participant suggested that a comprehensive balancing of benefits and risks must begin with a complete accounting of the political values, rights and principles at stake in data reuse scenarios. The rights and values explicitly discussed by workshop participants are described below.

### 1. Privacy

As discussed in the previous section, privacy is one concept that resisted a consensus definition and valuation. Some participants suggested that they would not feel as if their privacy had been violated merely because their personal data collected for one purpose was later reused for another purpose without their consent. One participant, suggesting a distinction reminiscent of differential privacy, proposed a definition of privacy that distinguished “facts about [or specific to] an individual” from “facts about the world.” A fact about an individual, the participant proposed, is something that could not be learned if the individual withholds their data. On the other hand, if the individual is excluded from a dataset, any information that could still be inferred about the excluded individual would be a fact about the world, and would not implicate the concept of privacy with respect to that individual. Such a conception of privacy may have broad implications for the appropriateness of restricting access to statistical datasets compiled using methods that satisfy the definition of differential privacy, which purports to provide a way to analyze datasets and permit the drawing of inferences that would be consistent regardless of whether any one individual’s data is actually included in the dataset. If this conception of privacy were generally accepted, and differential privacy accepted as a way to guarantee this sort of privacy protection, then any differentially private dataset could potentially be shared with far fewer restrictions on access. One participant, however, questioned whether this notion of privacy meshed with traditional notions of privacy, since under this model even data about a very small group of people could be said to no longer fall within the realm of privacy concerns. The participant noted that inferences about a particular individual based on information collected about similar people can still have a direct impact on an individual, raising the question of whether “collective privacy” should exist. This area has been flagged as a topic for a future roundtable to be convened by Georgetown Law, as discussed in the conclusion of this report, below.

### 2. Data Protection

One participant proposed that perhaps the notion of data protection—a term of art referring to the European Union’s FIPs-centric approach to privacy law—might provide a more concrete footing on which to ground the discussion. The European Union’s General Data Protection Directive (GDPR) avoids reference to privacy and focuses instead on ensuring that “natural persons” retain control over the collection, processing, and sharing of their personal data, and that data handlers employ

security measures sufficient to safeguard the confidentiality of personal data.<sup>14</sup> At least one critic of the ambiguity of *privacy* agreed that using data protection would clarify the conversation, but noted the FIPs and the concept of data protection do not deal with privacy exclusively, and that there are other, “broader social issues” at stake.<sup>15</sup>

One participant noted that, while a balancing framework that weighs privacy concerns against the function of data in society is perhaps a more functional approach, it leaves open the crucial broader question of whom can be trusted as the gatekeeper deciding when and how individuals’ data will be reused. Several participants raised concerns that strict data protection regulations governing the reuse of data risks limiting access to specific researchers or approved parties, in a way that will reinforce or systematize existing ills and inequities.

### 3. Addressing Bias & Inequality

Several participants noted that big data and predictive analytics have been shown to increase and exacerbate social inequality. These issues have been potentially made worse due to the “black box” nature of machine learning algorithms. For example, recent attempts at criminal justice reform have employed demographic sorting techniques that—while avoiding overt racial discrimination—ultimately resulted in the selection of proxies for race, such as residence in certain neighborhoods and certain elements of family or personal history that are closely related to historic racial discrimination.<sup>16</sup>

Some participants remarked how, because most machine learning algorithms highlight *correlation* without searching for underlying *causation*, machine learning analysis can be used in damaging ways that reinforce existing feedback loops without questioning the underlying data.<sup>17</sup> Related to some of the participants’ concerns regarding reinforcing historic inequality, if researchers only have access to data on certain sub-populations, such as the universe of all convicted criminals, instead of data on the entire population, then any inferences drawn from those limited datasets will be tainted by the historic racial and socio-economic biases of the criminal justice system. One participant expressed concern that operating purely off statistical correlations can, in some cases like medicine,

---

<sup>14</sup> *Data Protection in the EU*, [https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en). The full text of the GDPR is available at <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L.2016.119.01.0001.01.ENG&toc=OJ%3AL%3A2016%3A119%3ATOC>.

<sup>15</sup> For example, the GDPR recognizes that “The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality.” GDPR Paragraph 4.

<sup>16</sup> See, e.g., Julia Angwin & Jeff Larson, *Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*, PROPUBLICA, Dec. 30, 2016, <https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say>.

<sup>17</sup> This issue has been explored in depth by authors such as CATHY O’NEIL, *WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY* (Crown, 2016).



actually kill people, while in cases like sentencing reform, may create a much more opaquely racist system.

#### 4. Enabling & Promoting High-Quality Scientific Research

Several participants noted concerns about systematizing the unequal distribution of access to data through future regulation. For instance, if the regulatory framework erects or reinforces layers of redundancy and inefficiency as hurdles to access, in an attempt to restrict access by imposing barriers based on financial and other resource expenditures as prerequisites to meaningful access, then only large institutions like corporations may be capable of overcoming those obstacles, and citizen scientists would be foreclosed from access altogether. Several participants urged caution in adopting regulations that might compound existing inequality of access problems.

One participant suggested that an unflinching reliance on individualized consent could lead to similar problems regarding the quality of any outputs: “in most cases, if you want to do individual consent after the fact, you’re just preventing yourself from being able to answer the question altogether. If some small non-random subset denies access to their data, it won’t be useful if you can’t get to the whole universe.”


Another participant suggested that limiting access to and reuse of datasets could lead to similarly poor results: “without some sort of re-usable data assets, every time you ask [the holder of a dataset] for data . . . you’re going to get a different extract, and you will have just junk science.”

Together, these concerns make the question of *who* has access to data processing very important, but the argument can cut two ways. On the one hand, the complexities of machine learning (coupled with privacy concerns) may argue in favor of releasing data sets only to a very limited set of trusted third parties who ascribe to particular protocols for using the data and subject themselves to accountability mechanisms enforced by institutional gatekeepers. On the other hand, the call for greater algorithmic accountability may argue in favor of allowing a greater range of individuals to access particular data sets so that conclusions can be tested and challenged in the open air.

#### 5. First Amendment Rights & Values

When the question of whether to regulate the use of publicly-available data came up, several participants expressed serious concerns about the legality of such regulation in light of the First Amendment. As one participant remarked, short of making it illegal for people to do science, it would be impossible to prevent researchers from drawing upon otherwise publicly-available data. However, while there may be First Amendment barriers to preventing the mere use of such data in research, participants pointed out that harmful uses of the *results* of that research would still be amenable to regulation. For instance, restricting the ability of health insurance companies to set premiums based on newly available analytics probably would not implicate any First Amendment rights of the insurance companies, but prohibiting the research that produced the new analytics in

---



the first place could raise First Amendment issues. For example, one participant referred to the fact that many state regulators prohibited automobile insurers from using credit scores to set premiums, despite their utility as effective predictors of risk.

## 6. Serving Communities & People in Need

One of the most prominent potential benefits of big data and machine learning is the ability to more efficiently assist communities and people in need. For instance, one of the hypothetical scenarios used to frame the workshop discussion involved the potential of non-profit legal aid organizations to make use of their large amounts of case data to create an automated system that could draft pro se complaints based on an intake interview and alert clients to the likely steps required for them to seek relief in court. Some participants noted that this hypothetical scenario illustrates how the people who stand to benefit from data reuse are not the same people whose data is being reused, which further demonstrates the challenges of clearly distinguishing public and private benefits, and how to handle issues related to notice and choice.

One participant suggested that if instead it was a private law firm that wanted to use machine learning to improve outcomes for future clients, it seems unlikely that a client would allow the firm to use their data to then make other clients' cases more effective. The participant noted that, in fact, this is already happening absent any machine learning system— "the whole reason these law firms exist is to channel the expertise they cultivate by defending past clients." The participant noted that there is "an interesting question here about who has a claim to the benefits that can be derived from these models, and they are ultimately distributional questions about to whom the benefit accrues."

One participant commented that the same tool that may have been designed to help, for instance, tenants without lawyers to defend their rights against landlords, "could be flipped on its head and used by landlords . . . to disadvantage those whom you think you are helping. . . this [is a] problem, where you think you are advancing a social good, but there's a risk that that identical tool will be turned back against you."

Having outlined the challenges of identifying, defining and weighing the benefits and costs of data reuse—and acknowledging the important work that remains to be done in this area—the workshop participants proceeded to discuss prospects for regulating data reuse in the future, including an analysis of potential regulations on data collection and use, and the possible ways to assign responsibility for regulating access to previously collected datasets.

---

## PART III. REGULATING DATA USE IN THE MACHINE LEARNING WORLD

Workshop participants discussed three potential points of contention over data reuse going forward, revolving around questions of (1) whether and how data ought to be collected, consolidated, and organized; (2) whether and how the types of research entities and the uses to which data is put ought to factor into the decision of whether to allow access to datasets; and (3) who ought to be ultimately responsible for deciding when to allow the reuse of data.

### A. REGULATING COLLECTION & CONSOLIDATION OF DATA

One workshop participant suggested that there are two significant moments in the lifecycle of data which may provide opportunities ripe for regulation: (i) the moment when data is collected—when a data subject or data owner provides information to another party; and (ii) the moment when an opportunity arises to analyze or to reuse the data.


#### 1. Regulating Data Collection

Some workshop participants who spoke suggested that regulating the moment of data collection has been particularly challenging in the context of data reuse due to the difficulties involved in providing meaningful notice and choice. Whether information is collected by the government or by private entities, most people are likely to check the box on the form that acknowledges consent to the sharing and reuse of their data, in most cases because doing so is required in order to receive a government benefit or use a software application or other service. Participants noted that the collection phase may be ripe for additional rules or regulations limiting the types and/or amounts of data collected on individuals, according to what is appropriate for the provision of a particular service or program. Some participants suggested that additional transparency at the moment of collection may be needed. Other participants remarked that the tensions between the FIPs concepts of notice and consent, and data sharing and reuse, which many agreed rendered individual choice illusory and meaningless, indicate that while transparency is important, the moment of collection is not the most opportune moment for effective regulation.

#### 2. Regulating Data Consolidation: Data Pools vs. Data Lakes

Participants discussed the consolidation of different datasets collected at different times and often for different purposes, referring to the notions of “data pools” and “data lakes.” Currently, most data exist in “pools,” siloed off from other similar datasets. For example, hospitals and universities each maintain their own independent databases on their students and patients, and if researchers want to collect and consolidate data from multiple institutions, the process can be cumbersome, redundant, and resource-intensive. Participants debated whether these practical barriers to accessing consolidated “data lakes” might be beneficial and effective in controlling and restricting access to data.

---



One participant argued that even if a rule prohibited certain data lakes, data researchers could still do research. They would just face additional friction and burden every time they wanted to combine data. Such a rule might also require each subsequent researcher to jump through the same hoops to combine the same data, which might seem redundantly wasteful, but which the participant defended as a form of “desirable inefficiency.”

In response, one participant suggested that some may consider it unethical to erect or maintain artificial barriers or unnecessary inefficiencies between datasets. Doing so, the participant argued, could potentially result in delays in the development of life-saving scientific advancements. Another participant agreed, on the grounds that implementing arbitrary inefficiency would unfairly prejudice small research entities as compared to large corporate entities that have the financial resources to purchase access to restricted data sets.

There was greater consensus, however, on the idea that data lakes raise security concerns. Many participants seemed to concede that the creation of one centralized data lake across institutional datasets—at least for more sensitive types of data—would be inadvisable, due for example to risks of data breach jeopardizing much more data than is currently the case in most institutions. The recent breach of credit reporting agency Equifax was offered as one example of why such massive consolidations of data raise the stakes in terms of data security.

One participant referred to the Commission on Evidence-Based Policymaking’s solution to this issue.<sup>18</sup> Although the Commission’s Final Report advised against the creation of anything like a centralized data lake, the Report proposed the creation of a “National Security Data Service (NSDS), which is a sort of higher level centralization or higher level access point that can potentially give [researchers] access to a lot of these individual institutions rather than having to approach each agency and sub-agency.”<sup>19</sup>

Still, several participants urged that data lakes may be beneficial, at least in certain contexts. One participant noted that mandates requiring researchers to delete their datasets after completing their studies serve to perpetuate the “reproducibility crisis,” where it is difficult or impossible to retest and recreate findings and where research on siloed data sets will return disparate results that will not prove useful for society at large.

---

<sup>18</sup> The Commission on Evidence-Based Policymaking, established by the Evidence-Based Policymaking Commission Act of 2016 (P.L. 114-140), brought together 15 “individuals with experience as academic researchers, data experts, program administrators, and privacy experts,” who were charged with developing “a strategy for increasing the availability and use of data in order to build evidence about government programs, while protecting privacy and confidentiality.” COMMISSION ON EVIDENCE-BASED POLICYMAKING, ABOUT CEP, <https://www.cep.gov/about.html>.

<sup>19</sup> *The Promise of Evidence-Based Policymaking, Report of the Commission on Evidence-Based Policymaking*, Recommendation 2-2 (Sept. 2017), <https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf>.

## B. REGULATING USE & ACCESS TO DATA

Participants noted that the second moment, the moment of use or reuse, is currently handled very differently by different entities, and that there is a great deal of diversity in approach even among government agencies to regulating data reuse. In the federal government context, different agencies often have different legislative requirements governing their use and sharing of data, and those statutory requirements in turn have been interpreted by different agency lawyers. The ongoing process of interpreting and applying legislative mandates in light of new research and technology has increased the diversity in approaches even among federal statistical agencies, with some agencies like Census and IRS being more risk averse in their data sharing practices than others. One result is that the ultimate outcome of agency decisions on data sharing are indeterminate, and traditional canons of statutory interpretation are generally unhelpful in attempting to alleviate this indeterminacy.

### 1. Characterizing & Classifying Purposes & Uses of Big Data


Currently, agencies and institutions vary widely in their approach to data sharing and reuse based primarily upon the types or classes of data collected and the reuse for which the data is sought. At least one participant wanted to push back against the sectoral or “domain” categorization of data reuse, where for example “health” information is treated in a categorically different manner than “financial” information. The participant proposed a spectrum of *use* classifications, with emergencies on one end and the general collection of knowledge at the other. Emergency use cases, such as a particular public health or national security crisis requiring a particular deployment of data analysis, might prove appropriate for increased sharing and reuse, because the purpose is easily specifiable and therefore more easily evaluated under a risk-benefit balance analysis. Broader, non-particular purposes beyond the general collection of knowledge, however, might be more appropriate for heightened scrutiny and skepticism, because both the risks and benefits are potentially unknown and unknowable.

### 2. Privileged Access vs. Open Data

Another central point of contention among participants was the dichotomy between the frameworks of *privileged access* and *open data*. Some participants spoke out strongly in favor of a privileged access model, but while most participants at least seemed sympathetic to limited access in certain contexts, there was spirited debate over specific approaches to such a model.

At least one participant cautioned that a privileged access model, and especially one that includes artificial inefficiencies as barriers, could very well “make research by independent researchers with very tiny budgets impossible” in contrast with large companies. The participant contended, “part of our goal needs to be enabling research by independent researchers, at least in medicine.”

---



Under a truly open data regime, datasets might be posted on the Internet and available to the general public, to freely allow for citizen science research. Many participants bristled at the notion of such free and open access to large data collections. One participant even contended that allowing access for citizen science opens the data floodgates to “trolls and hackers,” arguing that there was little value and incredible risk in such open access. The participant instead argued for privileged access, regulated by strong data use agreements and regulatory controls.

One supporter of the privileged access model shared insights from her own experience implementing such a model:

*This idea of Privileged access makes a ton of sense. [At the speaker's organization], we can literally fire people if they misuse the data. And we've taken it one step further: we've now allowed a few external academics—a handful, I could count them on one hand—access to the data. We considered whether we should go further in opening that data up. But with each extra step you lose total control over turning the spigot off. If we keep access scarce, we have control over the quality of the scientists, and it's become a bidding exercise that considers the alignment of incentives, like will an employee lose their job if they're a bad actor, or will an outside researcher forfeit their academic reputation or the academic value of being able to access this data set that no one else has access to? It has actually worked quite well.*

One participant urged consideration of a sliding scale or spectrum, where the degree and type of restrictions on access should scale with the sensitivity of the data. Many participants who spoke seemed to agree that a sliding scale or a hybrid-access model would be necessary, as most who spoke seemed to agree that some types of data would be too risky, or sensitive, to release to the public. Despite some dissent, most participants agreed that there were at least some situations where privileged access made sense, and the debate proceeded to questions regarding how and to what extent access should be limited.

Moreover, one participant noted, even purportedly open repositories have a habit of finding their way into closed repositories, often of private for-profit entities who have an interest in restricting access to data or keeping findings from research secret. Corporations who lock down what they perceive as proprietary data perpetuate unequal distributions of access and trade secrecy may make, as one participant noted, “trolls and hackers of us all”—that is, at least, of anyone else wishing to use that data.

Finally, at least one participant who generally favored open data but expressed sympathy for a privileged model, noted that such a privileged access approach would work a lot better if the person or entity in charge of granting and denying access sat atop an enormous data lake, in order to avoid the aforementioned reproducibility crisis said to result from meting out access piecemeal.

---

Discussion then moved to the question of who exactly should wield the ultimate responsibility of regulating access to such vast and valuable data sets.

### C. ASSIGNING RESPONSIBILITY FOR REGULATING DATA REUSE

One participant referred to the ad-hoc, piecemeal, and agency-by-agency/institution-by-institution approach to controlling access to data as regulation by “Philosopher Kings,” wherein individual data privacy experts and in-house counselors act as the ultimate gatekeepers of their entity’s data pools. At least one participant who spoke to this point lamented that the current so-called Philosopher-King model of regulation was far too vulnerable to subjective value judgments and gatekeeper bias.

Several participants who spoke seemed to agree about the need for greater consistency in institutions’ approaches to allowing access. One participant suggested that instead of delegating the task of interpreting statutory obligations to agency lawyers pursuant to the traditional approach to administrative procedure, a more agile, responsive, and structured regulatory framework is needed. Still, the questions of what those processes ought to be and what the standards and criteria ought to look like for granting or denying access, were left largely unresolved.


Working from the premise that some form of privileged access model is necessary to protect the types of information society deems most sensitive, participants considered three general approaches to privileged access: (a) developing a framework similar to the “Belmont principles,” which typically govern research involving human subjects, and thereby creating a mechanism analogous to institutional review boards (IRBs); (b) strengthening and expanding upon traditional legal mechanisms for regulating data pooling and reuse, such as the Fair Credit Reporting Act (FCRA) or Health Insurance Portability and Accountability Act (HIPAA); and (c) a so-called “trusted entities model,” for particular data repositories, building off the conceptual framework of credit reporting agencies (CRAs).

#### 1. Belmont-Plus & IRB-Light Models

Participants discussed the viability of expanding application of the Belmont principles<sup>20</sup> and implementing some form of institutional review board (IRB) for data-holding organizations to better govern the sharing and reuse of data in their possession. While participants noted that the

---

<sup>20</sup> The Belmont principles are a set of basic ethical principles, summarized in the Belmont Report and first published in the Federal Register in 1979, which were identified by the National Commission for the Protection of Human Subjects of Biomedical & Behavioral Research pursuant to one of the charges of the Commission’s enacting legislation, the National Research Act (Pub. L. 93-348). Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 44 Fed. Reg. 76, 23192 (Apr. 18, 1979).



Belmont-IRB model suffered from shortcomings similar to those found in FIPs—namely, that its focus on individual consent creates the same problem of implementation when it comes to re-collecting and re-purposing data after the fact—some potential adaptations of the traditional Belmont-IRB model were discussed favorably, even by participants who were generally opposed to stricter privileged access frameworks.

One participant, while expressing support for the inquiry into whether some modified Belmont-IRB model might be extendable to other use cases, also highlighted some reasons for doubting the model's appropriateness for the machine-learning context:

*The concept [of human subjects research] is that we are doing some sort of systematic inquiry involving living human beings to generate generalizable knowledge. . . . I am curious whether there is a way to apply those same principles to the particular uses we've been talking about, [such as] uses targeting specific individuals, or uses related to profit, or improving models that are proprietary, all of which is not generalizable knowledge in the sense that an IRB would define it in the context of human subjects research. . . . Or whether the nature of machine learning itself demands a whole new set of principles and infrastructure for those separate cases.*

#### *a. Belmont-Plus: Belmont for All Society*

One participant proposed that a formal infrastructure implementing and applying the principles outlined in the Belmont Report might provide a framework for regulating access to data. The Belmont Report broadly identified three ethical principles for research involving human subjects: (a) respect for persons; (b) beneficence—i.e., benefit-maximization and avoidance and minimization of harm or risks of harm; and (c) justice—i.e., fair and equal administration and distribution of costs and benefits, and obligation to share research findings.

##### *i. Respect for Persons*

First, participants noted that the principle of respect for persons closely overlaps with the notice and choice elements of the FIPs by requiring individualized and informed consent from data subjects. One participant pointed out that reliance on Belmont principles moves the conversation away from regulation of access based on the purposes to which the data is put, and back to a framework designed to protect individuals and place ultimate control over participation in the hands of the data subjects—a framework several participants had discounted as overly cumbersome and impractical in the earlier discussion of the future viability of the FIPs. One participant suggested that if individualized consent is de-emphasized, the “respect for persons” principle could be construed in conjunction with the other two principles, beneficence and justice, to adapt the Belmont principles to the big data and machine learning context.

---

## ii. Beneficence

The viability of the second principle, beneficence, was also questioned by participants who recalled the earlier-discussed challenges of balancing benefits against risks and harms; namely, the difficulty of identifying such benefits and harms at the outset due to the dynamic and unpredictable nature of machine learning systems. If Belmont principles were adapted, with the focus cast not on individuals but on society at large, and oversight concerned not with *a priori* evaluations of benefits and risks but with the imposition of constraints on harmful outcomes, at least one participant suggested, perhaps a Belmont-for-Society type framework might prove more practicable.

## iii. Justice

Finally, several participants who spoke were sympathetic with the import of the third Belmont principle of justice. If the other two principles are interpreted through the lens of “justice”, the framework might be adaptable by shifting the focus of oversight to ensuring that no person or class of people bear an undue or disproportionate amount of harm or risk of harm, and that the benefits of data reuse do not disproportionately accrue to one privileged entity or group of entities, but rather to society at large. More than one participant who spoke seemed to suggest that one way this might be implemented, in line with the justice principle’s call for equitable distribution of benefits, would be to require that researchers share their findings.

Ultimately, many of the participants who spoke out in favor of the privileged access model also supported some adaptation of Belmont-type principles, expanding on the well-established norm that researchers should not be able to collect and use information about a person’s body without substantial oversight and restrictions. To that end, participants turned the discussion to the concept of IRBs (institutional review boards), which were developed in response to the recommendations of the Belmont Report, in order to implement a system of upholding the principles enshrined in the Report.<sup>21</sup>


### b. IRB-Light for Private Entities

Workshop participants debated the prospect of requiring private entities to establish IRBs for data sharing and reuse, although perhaps with less stringent standards than those of traditional IRBs in the context of hospitals and similar institutions.<sup>22</sup> One participant mentioned that Facebook now

---

<sup>21</sup> See Jennifer M. Sims, A Brief Review of the Belmont Report (2010).

<sup>22</sup> For example: “Another possibility is that for-profit entities want to make their information available for certain kinds of knowledge creation that would be helpful; for example, a bank trying to help figure out how to make mortgages available to different people. If you then allow the data to go with data scientists under a set of careful rules that limit the risks of re-identification, that mask the names, that create information that’s useful, you’ll get lot of benefit and very little risk because you have a lot of controls at the technical and administrative level—and hallelujah, we get some knowledge with very little risk. That’s a model of . . . privileged access, and IRBs



claims to have something like an IRB, which the company actually refers to as an IRB, but which bears little resemblance to the rigor and regimentation of hospital IRBs.<sup>23</sup> Another participant was adamant that, notwithstanding what Facebook might want to call its process, and notwithstanding its lack of rigor and regimentation, it is *not* an IRB, in many respects.

Another participant suggested that in practice, at least in their experience, IRBs often offer little more than the veneer of protection.

One participant cautioned that, similar to the FIPs, reliance on Belmont principles may place too much responsibility in the hands of individual subjects:

*[Belmont principles] are designed to protect research subjects, they are not a mechanism to evaluate whether the research itself is a good thing to pursue. And it's not obvious to me we have any mechanism to evaluate the value of research. What happens instead may be that, incidentally, we expect that by soliciting consent from participants, they act as the adjudicator of whether this is an acceptable research question to pursue. If you persuade the participant, beyond the fact that you'll guarantee their individual privacy, this is actually a research project worth pursuing, or it's going to generate knowledge that the participant concludes is worthwhile enough that they are willing to contribute. I don't know if that is the right mechanism, it seems more like an accident in the way we've arranged it. There may be a much better way assessing whether this is something we should use this data for.*

Another participant pushed back on this characterization of the Belmont principles, noting that “there is a provision in there that asks how valuable is the research, and they try to balance that against potential risks to the data subjects.” However, the participant noted, this consideration brings the conversation back one of the central differences of machine learning, which is that “often, the real use of the information is not something you will know before you do a lot of analysis of it, and so you really need a process where you vet the quality of the people who are doing the research and the process they are going to be going through in order to keep it safe and secure.”

Another participant also raised the question of who should sit on IRBs, and whether it would be feasible or even possible to require that outside stakeholders and other people representing the

---

should be talked about more, because mostly what the regulatory system will be is IRBs everywhere internally. Because that's mostly all you can do, and people will get outraged if you violate it.”

<sup>23</sup> See Zoltan Boka, Opinion, *Facebook's Research Ethics Board Needs to Stay Far Away from Facebook*, WIRED, June 23, 2016, <https://www.wired.com/2016/06/facebook-research-ethics-board-needs-stay-far-away-facebook/>.

---

general interests of data subjects have seats on the IRBs of private entities. This question was not discussed at length and is possibly a fertile ground for future discussion.

## 2. Adapting Traditional Legal Mechanisms

Workshop participants also discussed the adaptability of traditional legal mechanisms to regulating data reuse in a machine learning context. In particular, participants mentioned legal protections arising from contract law and intellectual property law, including license limitations and antitrust regulations.

Participants first discussed the possibility that contractual sharing and use agreements could continue to serve an important regulatory role. If the proposal for some form of centralized data repositories were pursued, some participants suggested, strong contractual conditions on access to the repositories would be essential. One participant cited relevant research by Katherine Strandburg, Brett Frischmann, and Michael Madison into frameworks for information and resource-pooling arrangements, specifically referencing their co-authored article *Constructing Commons in the Cultural Environment*.<sup>24</sup> The participant noted that these data commons could have a hybrid-access/IRB flavor, reflecting previous comments by several of the workshop participants.


One participant noted that the reliance on legal limitations sounding in property and contract raises a further question regarding the handling of inevitable abuse of the contractual relationships. For instance, although participants had discussed examples such as credit reporting where interests and incentives exist to share data among competitors, the participant pointed out the likelihood that in many contexts arising in the future, one giant company—Ali Baba, for example—may have access to all the useful data necessary to achieve their goals, and they would thus have no incentive to share that data with other companies. The participant wondered whether this is the sort of scenario where the law ought to provide some sort of override button to compel even private entities to share and pool their data. Antitrust provisions were one relevant traditional legal measure discussed in the *Constructing Commons* article.

## 3. Trusted Entities Model

Continuing the discussion of potential approaches to regulating consolidated data lakes, workshop participants considered whether trust could be placed in certain entities to maintain such data lakes and regulate access to their information by acting as data clearinghouses. Participants noted that the idea of data clearinghouses is attractive not only because they would provide a more consistent

---

<sup>24</sup> Michael J. Madison, Brett M. Frischmann, & Katherine J. Strandburg, *Constructing Commons in the Cultural Environment*, 95 CORNELL L. REV. 657 (2010). The authors discuss potential governance mechanisms of the commons, such as membership rules, resource contribution or extraction standards and requirements, conflict resolution mechanisms, and sanctions for rules violations.



and regimented regulatory framework, but because of the promise of uniform and centralized control over access. The opening up of data sets for analysis comes at the expense of control over the use of such data. A trusted entity could better control this dataflow by overseeing access in a methodical way.

One participant asked whether the Fair Credit Reporting Act's (FCRA) model of centralized credit reporting agencies (CRAs) and strict regulations on access to their data based on approved uses might be expandable to other industries. As one participant pointed out, the FCRA-CRA model is analogous to the data reuse scenarios discussed by the workshop, in that the individual data subjects have little or no say over whether their data is included in the database. Recognizing individuals' loss of control and the highly-sensitive nature of such data, Congress enacted specialized protections that could be appropriately expanded to other scenarios.

Several of those who spoke, however, seemed to agree that reference to the FCRA-CRA model actually illustrated the dangers of centralized clearinghouses, as recent data security breaches demonstrate the risk of creating vast lakes of sensitive information.<sup>25</sup> One participant noted that the Commission on Evidence-Based Policymaking rejected the creation of a single data clearinghouse (see *supra* at 14), in part for this reason.

Another participant noted that discussion of a trusted entities model highlights the existence of an institutional gap. Such a model would require a central repository for providing access, but no such infrastructure currently exists, and building it would require massive technical overhead. The participant suggested that such a resource-intensive undertaking may not be practically feasible, because it is unlikely that there would ever be one single project that would justify such expenditures. The participant suggested that new classes of institutions might be required in order to ever implement a trusted entities model.

---

<sup>25</sup> See, e.g., FTC, Consumer Information, *The Equifax Data Breach: What to Do*, Sept. 8, 2017 ("If you have a credit report, there's a good chance that you're one of the 143 million American consumers whose sensitive personal information was exposed in a data breach at Equifax."); see also *supra* Part III.A.2 discussing risks inherent in creating consolidated data lakes and possible alternative approaches discussed by the Commission on Evidence-Based Policymaking.

---

## PART IV. TOPICS FOR FUTURE DISCUSSION

### A. IS DIFFERENTIAL PRIVACY “PRIVACY”?

One recurring theme throughout the workshop centered around the lack of consensus on how to define and conceptualize privacy: an essential question as we determine what interests society should work to protect. At least one participant suggested that differential privacy<sup>26</sup> offers not only a comprehensive and effective approach to protecting privacy, but also a way to formally define privacy itself.<sup>27</sup> Georgetown’s Professor Paul Ohm, the convener of the workshop, proposes that the next workshop in this series will likely center around the question of whether differential privacy can deliver on these promises, and will seek to answer whether differential privacy is, in fact, “privacy”—in other words, whether it is sufficiently correlated with the interests privacy seeks to protect.

### C. OPEN DATA, PRIVILEGED ACCESS, & DATA CLEARINGHOUSES

Most participants seemed to agree that neither a purely open data environment nor an entirely closed access model would suffice as a one-size-fits-all approach to protecting data in the era of machine learning. Instead, many of the participants who spoke expressed sympathy for a hybrid-access model. However, many questions remain about just what such a hybrid-access approach will look like in practice, and how to get from our current system to a framework that is more responsive to the practical realities of big data and machine learning.

Of the many open questions at the end of the workshop, some of the most contentious were:

1. Is the creation of large data clearinghouses even feasible, and if so, in what contexts would they be worthwhile?
2. If data remains in smaller silos or data pools, what sort of regime would best facilitate research?
3. How much care should be taken to ensure that access controls do not disproportionately restrict the ability of smaller research entities or allow large corporations to monopolize research in the future?
4. Who should be in charge of regulating access to data?
5. Should there be a movement toward model principles guiding how individual organizations manage access to data—and if not model principles, at least a call for greater transparency around how organizations currently approach this question?

---

<sup>26</sup> See Dwork, *Differential Privacy*, *supra* n.13 at 1–12.

<sup>27</sup> See *supra* Part II.B.1, discussing the concept of differential privacy.



### C. IDENTIFYING & WEIGHING BENEFITS & RISKS

Finally, several participants repeatedly stressed the inherent difficulties in identifying and balancing the benefits and risks of data reuse in a machine learning era. At the workshop's conclusion, it remained unclear whether such analyses are even possible given the dynamic and far-reaching nature of machine learning techniques. The workshop participants' attempts to operate under traditional notions of public and private benefits and risks, and to identify those benefits at the outset of any research project as traditionally required under the FIPs, highlighted what may be the most substantial differences.

## PARTICIPANTS

Katherine Alteneder

Alvaro Bedoya

Miranda Bogen

Cindy Chance

Anupam Chander

Aloni Cohen

Julie Cohen

Elizabeth Edenberg

Nico van Eijk

Aaron Fluitt

Jonathan Frankle

Alexandra Givens

Fiona Greig

Janis Kestenbaum

Yafit Lev-Aretz

Maggie Little

Mark MacCarthy

Mark Maloof

Carlos Manjarrez

Laura Moy

Adam Neufeld

Mark O'Brien

Amy O'Hara

Paul Ohm

Joel Reidenberg

Aaron Rieke

Tanina Rostain

Aaron Roth

Cameron Russell

Josh Swamidass

Peter Swire

Ali Whitmer

Self-Represented Litigation Network

Georgetown Law | Center on Privacy & Technology

Upturn

Georgetown University | Kennedy Institute of Ethics

Georgetown Law

MIT

Georgetown Law

Georgetown University | Kennedy Institute of Ethics

University of Amsterdam

Georgetown Law | Institute for Tech Law & Policy

MIT

Georgetown Law | Institute for Tech Law & Policy

JP Morgan Chase Institute

Perkins Coie

NYU Information Law Institute

Georgetown University | Kennedy Institute of Ethics

Software & Information Industry Association

Georgetown University | Dep't of Computer Science

Legal Services Corporation

Georgetown Law | Center on Privacy & Technology

Georgetown Law | Institute for Tech Law & Policy

ProBono Net

Stanford Institute for Economic Policy Research

Georgetown Law

Fordham Law

Upturn

Georgetown Law

U. Penn.

Fordham Law | Center on Law & Information Policy

Washington University in St. Louis

Georgia Tech

Georgetown University

## ABOUT THE CONVENOR

### GEORGETOWN INSTITUTE FOR TECHNOLOGY LAW & POLICY

The Institute for Technology Law & Policy at Georgetown Law is dedicated to training the next generation of lawyers and policymakers with deep, practical expertise in technology law and policy, and providing a forum to address pressing problems and opportunities involving technology and the law.

Among its initiatives, the Institute seeks to bring together policymakers, academics, business leaders and technologists to develop policy solutions informed by a deep understanding of technology. It also works to promote tech competency among policymakers, hosting an annual two-day immersion program for congressional staff to learn about new developments in technology policy, and workshops to support legislative staff and other experts. For more information visit [www.georgetowntech.org](http://www.georgetowntech.org).

The Tech Institute expresses its sincere thanks to Axa Research Fund for supporting this event.

---



Georgetown Institute for Technology Law & Policy  
600 New Jersey Avenue NW | Gewirz 309  
Washington, D.C. 20009  
Email: [TechInstitute@law.georgetown.edu](mailto:TechInstitute@law.georgetown.edu)  
[www.georgetowntech.org](http://www.georgetowntech.org)

---