Privacy Protection in the Era of "Big Data": Response to Office of Privacy Commissioner's Discussion Paper on "Consent and Privacy"

Colin J. Bennett,
Department of Political Science,
University of Victoria,
BC, Canada
www.colinbennett.ca
cjb@uvic.ca

(This response is adapted from "Privacy Protection in the Era of 'Big Data': Regulatory Challenges and Social Assessments," in Bart van der Sloot, Dennis Broeders and Erik Schrijvers, *Exploring the Boundaries of Big Data* (Amsterdam: Amsterdam University Press, 2015), pp. 205-227, with Robin M. Bayley.

The discussion paper "Consent and Privacy" raises a number of issues about the "consent model" upon which PIPEDA is based. Pressure on this model is more acute than it has ever been, as a result of what some have termed the "Big Data" revolution. I would like to offer some reflections on the Big Data revolution in response to the analysis in the paper (pp. 6-7). We need to separate the reality from the hype in order to properly understand the nature of the challenge to the traditional consent model.

My overall position on the issues raised in the paper is that: 1) consent is, and should remain, the cornerstone of any privacy-protection policy; and 2) the system requires the implementation of all potential privacy instruments in the toolbox (regulatory, self-regulatory and technological). That was the central message of my 2006 co-authored book, "The Governance of Privacy" (Bennett and Raab, 2006). It is obvious that Canadian privacy protection needs: better and more transparent privacy policies that work across services; privacy by default and by design; better de-identification and standards for de-identification; the encouragement of privacy management frameworks and accountability; and better use of codes of practice, technical standards and privacy trustmarks. All are necessary, and none is sufficient. I am also persuaded that the Commissioner needs stronger enforcement powers.

For this response, however, I would like to focus on the questions addressed under the section on Ethical Assessments (pp. 22-24), and consider in what ways we might enhance and broaden privacy impact assessments to embrace the wider range of risks produced by Big Data analytics. The response is adapted from a recent chapter produced for a report in the Netherlands (Bennett and Bayley, 2016).

A considerable literature already exists on PIAs and on their development and implementation in different countries (Wright and de Hert, 2012). They are now institutionalized under many data protection regimes and will become, in some contexts, mandatory under the new EU General Data Protection Regulation (GDPR) (EU, 2016). In the main, however, these methodologies were developed before the challenges posed by big data analytics, and tend not to incorporate assessments of the broader discriminatory impacts of these practices.

Do existing PIA methodologies need to be revised to enable the evaluation of risk in the context of big data? What tools might be developed to assess the broader social risks of excessive surveillance and categorical discrimination? There are proposals for Surveillance Impact Assessments (Wright and Raab, 2012), and for more unified ethical frameworks, developed to guide data scientists (IAF, 2014; 2015). The paper questions whether the integration of existing PIA methodologies into a broader ethical frame is a critical condition for the mitigation of individual and social risks in this new era of big data analytics.

Finally, what other regulatory solutions have been proposed, both now and in the past, that could offer ways to allow the promise of big data analytics, and at the same time, to protect individual privacy rights? Is it really necessary to give up on the central tenet of

privacy protection law and philosophy in order to permit big data analytics to realize their potential? I do not think so. On the contrary, I would argue that the current debate tends to rest on a false dichotomy and a fundamental misunderstanding about the theory of information privacy that developed 40 years ago, and the data protection policies that it generated (Bennett, 1992).

What is Big Data, and is it a "Revolution"?

It is commonly assumed that the phenomenal and rapid expansion in the capacities of computing technology has entailed fundamental and qualitative changes in the "volume, variety and velocity" of data processing (US, Executive Office of the President, 2014). We are told that we live in a "data driven society", in which ubiquitous data collection from a bewildering variety of observational technologies is fundamentally changing organizational life and human values in revolutionary ways. As the US Executive Office of the President concluded in a recent report (2014, p. 54):

> Whether born analog or digital, data is being reused and combined with other data in ways never before thought possible, including for uses that go beyond the intent motivating initial collection. The potential future value of data is driving a digital land grab, shifting the priorities of organizations to collect and harness as much data as possible. Companies are now constantly looking at what kind of data they have and what data they need in order to maximize their market position. In a world where the cost of data storage has plummeted and future innovation remains unpredictable, the logic of collecting as much data as possible is strong.

For all the hype around the "big data revolution" we have to remember that the last fifty years have ushered in numerous claims about the revolutionary nature and potential of new technologies. If we can be persuaded that revolutions are occurring, we can also be persuaded to jump on board for fear of losing economic advantage or social esteem. "Revolutions" don't just emerge, they are constructed. And more often than not, these claims make simplistic assumptions about the trajectories of technological development, and gloss over complex social, political and economic assumptions. The messy, contradictory and ambiguous character of technological change is quite often simplified in the rush to encapsulate the present and extrapolate the future within a catchy phrase.

"Big data" is a socio-technical phenomenon that rests on a good deal of mythology (boyd and Crawford, 2012). It is not a precise scientific concept, but a highly contested idea that means different things depending on who is talking about it. There is, and will never be, any consensus on what "big data" means, nor on how its processing differs from the data analytical techniques of the past. There is no clear threshold at which point "data" becomes "big data." It is a highly fashionable, and therefore inherently suspect, idea that encompasses a complex array of technologies, practices and interests. "Big data" in and of itself means nothing, and signifies nothing, in the absence of a wider understanding of the organizations that are conducting the analysis, and an assessment of those organizations wider interests and motives.

The fundamental epistemology of big data is inductive, where data analysis is conducted without the benefit of a guiding hypothesis. The data itself, according to Mayer-Schönberger and Cukier, reveals the patterns; the human reasoning and judgment about causation comes afterwards (2013, ch. 4). Deductive methods, according to these authors, constrain the power of the data to find patterns that would not otherwise be discovered. They offer several illustrations (2013, p. 55):

> No longer do we necessarily require a valid substantive hypothesis about a phenomenon to begin to understand the world. Thus, we don't have to develop a notion about what terms people search for when and where the flu spreads. We don't need to have an inkling of how airlines price their tickets. We don't need to care about the culinary tastes of Walmart shoppers. Instead we can subject big data to correlation analysis and let it tell us what search queries are the best proxies for the flu, whether an airfare is likely to soar, or what anxious families want to nibble on during a storm. In place of the hypothesis-driven approach, we can use a data-driven one. Our results may be less biased and more accurate, and we will almost certainly get them much faster.

These arguments are, of course, controversial. They raise the possibilities of "spurious correlations" and about whether the data are reliable and valid proxies for the phenomena in question. They also rest on some questionable deterministic assumptions about the power of technology.

Mayer-Schönberger and Cukier also recognize that there is a "dark side to big data" which poses significant challenges to conventional legal instruments of privacy protection (2013, p. 170). Big data is a surveillance tool, and magnifies the capacity of organizations to monitor individuals' lives, to erode anonymity, and to discriminate on the basis of gender, race, age, location and other factors. Big data analytics using secret algorithms can lead to automatic and discriminatory judgments with widespread implications for the types of people most likely to engage in certain more risky behaviors (Pasquale, 2015). The "social-sorting" of the population using new technologies has been a theme in the surveillance literature for some time (Gandy, 1993, 2009; Lyon, 2003). These new tools permit a surveillance of the population in ways unimaginable a few years ago.

For instance, big data correlations have learned that workers with longer commutes quit their jobs sooner. Is it then fair to turn away job applicants with longer commutes? And what if those applicants tend to be disproportionately from minority populations (Robinson and Yu, 2014)? And is it appropriate for a company to assign you a credit score based on where you live, and inferences about the creditworthiness of your neighbors (National Consumer Law Center, 2014)? And is it acceptable for political parties to draw conclusions about how you might vote, on the basis of analysis of Facebook friends, twitter followers and other evidence of the "social graph"? (Bennett, 2013).

"Big data" leads to scoring practices, which are compiled based on financial, demographic, ethnic, racial, health, social, consumer and other data to characterize individuals or predict behaviors like spending, health, fraud, academic performance, or employability. Scores can be correct, or they can be inaccurate or misleading, but they are rarely transparent. Persons affected may not be aware of the existence of the score itself, its uses, and the underlying sources (World Privacy Forum, 2014). Citizens are generally unable to challenge the score, determine how the score is constructed, correct the data on which it is based, or opt out of being scored altogether.

The use of predictive analytics based on our online connections and activities can also inhibit freedom of association, and chill our online interactions. As the Electronic Information Privacy Center (EPIC) has stated: "The use of our associations in predictive analytics to make decisions that have a negative impact on individuals directly inhibits freedom of association. It chills online interaction and participation when those very acts and the associations they reveal could be used to deny an individual a job or flag an individual for additional screening at an airport because of the determination of an opaque algorithm, that may consider a person's race, nationality, or political views" (EPIC, 2014).

Where, then, do these trends leave personal information rights, and the many policy instruments that have been designed to protect those rights? What are the problems involved with the current information privacy or data protection model with regard to the regulation of big data? Some have claimed that the traditional privacy protection model, based on notice and consent, is now obsolete and counter-productive, and have insisted that the focus of regulation should shift to the "accountable" uses of personal data (Mundie, 2014).

Big Data and the Challenges to the Fair Information Principles Model

All privacy protection laws require the transparent communication of the purposes for which personal data will be processed. That transparency establishes a relationship of *trust* that personal data will not be re-used, re-purposed and disclosed to other organizations. This principle is at the heart of the theory of information privacy, and reinforces powerful social norms. And it governs both the processing of personal data, *and* its collection and capture.

There are three general and overlapping aspects of what we will call the "Fair Information Principles" model that, critics argue, are fundamentally challenged by big data and its implications. The first relates to the definition of personally identifiable information (PII) itself. Regulation in this area is triggered by the capture and processing of data that, in some way, relates to an identifiable individual (Schwartz and Solove, 2011). Increasingly the line between what is personal and non-personal data is increasingly difficult to draw for several reasons.

Personal data can more easily be re-identified from the combination of data elements

which, on their own, say little or nothing of about any one particular person. Our online tracks are tied to smartphones or personal computers through Unique Devise Identifiers (UDIDs), IP addresses, "fingerprinting" and other means. Given how closely these personal communication devices are associated with the individuals who use them, information linked to these devices is, to all intents and purposes, linked to individuals. The sophistication of contemporary re-identification science gives a false sense that data can ever be stripped of identifying markers (Ohm, 2010). And big data can increase the risk of re-identification, and in some cases, inadvertently re-identify large swaths of de-identified data all at once.

The problem is magnified in the context of the Internet of Things, where inferences about our behaviors and actions can more easily be drawn from the capture of data from objects in our possession – our phones, cars, household appliances and so on. Generally speaking, ordinary people go about their lives in complete ignorance of the technical identifiers that attach to these devices and constantly emit information about their personal lives. Further, decisions about the individual are increasingly made on the basis of inferences that are drawn about the categorical group to which we are presumed to belong. The world of big data feeds off this growing ambiguity about what is, and what is not, personally identifiable information (The New Transparency, 2014, pp. 71-85).

A second, and related, challenge is to the principle of "data minimization." Organizations are required to limit the collection of personal data to that which is necessary to achieve their legitimate purposes, and to delete that which does not conform to those purposes. The business model of big data is antithetical to these principles. Rather it incentivizes the direct and indirect capture and retention of any data, by any technical means. And whereas it was once cheaper to delete information than to retain it, the obverse is now the case (Mayer-Schönberger, 2011). "Data minimization is simply no longer the market norm" (Tene and Polonetsky, 2013, p. 260).

The final challenge relates to the clear definition and transparent communication about the purposes to which personal data are being processed. Some have argued that big data analytics require that presumption to be discarded, or at least fundamentally rewritten. The inductive power of analytics presumes that new purposes will and should be found for personal data, if the promise of the technology is to be realized. Mayer-Schönberger and Cukier (2013, p. 173) are emphatic on this point: "In the era of big data, however, when much of data's value is in secondary uses that may have been unimagined when the data was collected, such a mechanism to ensure privacy is no longer suitable." In the words of the US President's Council of Advisors for Science and Technology: "The notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data" (2014, p. 36).

Scott Taylor of Hewlett Packard, and a key participant in the Information Accountability Foundation, has likened the process of big data analytics to a chemical reaction (Taylor, 2014). Just as something new, and in some cases unpredictable, is created from the reaction of two chemicals, the same is true for big data. And just as one would not prevent the collection of certain chemicals because they have a chance, in reaction with others, to cause an explosion, he argues that we should not prevent the capture and

processing of personal data, because it might, when applied, have adverse consequences for individuals, groups and for society as a whole.

According to this position, there is no category of data that is *a priori* "none of your business." The promise of big data, we are told, assumes further uses for purposes not originally conceived. Craig Mundie, one of the members of the President's Council, (2014) insists that we should "focus on data use, not data collection." He adds a familiar argument that there is already so much personal data "out there" that it cannot be retrieved, and it is practically impossible to provide notice and seek consent for every conceivable use. Mundie, and others, envision a revision of the privacy framework that permits almost unlimited collection, in return for stronger accountability mechanisms that govern uses and disclosures. The argument broadly comports with a more long-standing effort to reorient privacy protection away from notice and consent, and towards an emphasis on accountability of organizational practices (Weitzner, 2008; Center for Information Policy Leadership, 2009).

Against this more "pragmatic" approach to privacy regulation, certain privacy advocates have weighed in to defend traditional interpretations of the privacy principles. Hoofnagle (2014), for instance, warns that "use-regulation advocates are actually arguing for a broad deregulation of information privacy." It amounts to the following: "1) Data companies can collect anything they want and analyze it however they please; 2) They are liable only for misuses of data, which businesses define themselves, narrowly; 3) If pressed, they can argue that use restrictions are unconstitutional censorship; and 4) Companies can purposely engage in those misuses, and only be liable when it causes concrete injury." Privacy pragmatism masks a "radical deregulatory agenda", according to Hoofnagle: "A regime that only pays attention to use erects a Potemkin Village of privacy. From a distance, it looks sound. But living within it we will find no shelter from the sun or rain." The Electronic Privacy Information Center (EPIC, 2014) also warned of huge dangers posed by data breaches and of the failure of organizations to adequately safeguard the personal data under their control, when those data are increasingly stored in larger and larger data repositories.

These debates will no doubt continue in the United States. One key, for countries like Canada, which after all can regulate Big data within the context of a comprehensive framework law, is to make a better use of more proactive risk assessments, and a broadening of PIAs into more comprehensive evaluations of the wider set of social harms.

Big Data and Privacy Impact Assessments (PIAs)

PIAs have been advocated to mitigate a range of organizational risks: vulnerabilities to organizational systems and assets; threats from malicious attacks; negative media publicity; loss of consumer confidence; infringement of laws; financial losses; dilution of brand, reputation and image; and so on (Wright and De Hert, 2012, pp. 14-15). They also arguably produce many positive benefits in their engagements with customers, stakeholders, regulators and others. They may operate as a learning experience for the

organization and its employees about what personal data the organization has, why it was collected, how it is stored, and to whom it will be communicated (Wright and de Hart, 2012, pp. 16-17). The prospective analysis of privacy impacts is now regarded as one critical element of good privacy management and governance.

Crucially, therefore, PIAs need to offer an identification of privacy risks *before* systems and programmes are put in place. PIAs are only valuable if they have, and are perceived to have, the potential to alter proposed initiatives in order to mitigate privacy risks. They also have to consider privacy risks in a wider framework, which takes into account the broader set of community values and expectations about privacy. PIAs should also be more than are more than the end-product or statement. They refer to an entire process, and appear to be more effective when they are part of a system of incentives, sanctions and review, and/or where they are embedded in existing project workflows or quality assurance processes (Warren et al. 2008).

However, PIAs vary across a number of dimensions: the levels of prescription, the application, the circumstances that might trigger PIAs, the breadth of the PIA exercise, the agents who conduct PIAs, the timing, the process or review and approval and the level of public accountability and transparency. In most jurisdictions where law or policy require or highly recommend that PIAs be conducted, an official PIA template, format or other tool to describe how they should be conducted, is provided. However, there is no simple formula for the conduct of a PIA. Each PIA should be dictated by the specific institutional, technological, and programmatic context of the initiative in question. Any PIA needs to be sensitive to a number or crucial variables: the size of the organisation; the sensitivity of the personal data; the forms of risk; and the intrusiveness of the technology (Warren et al. 2008).

The current reality, however, is that many PIAs are simply legal compliance checks, are not published, and certainly are not conducted with broad input from relevant stakeholders. Where they are conducted in a mechanical fashion for the purposes of satisfying a legislative or bureaucratic requirement, they are often regarded as exercises in legitimation rather than risk assessment (Warren et al. 2008).

To the extent that PIAs are seen as valuable tools by data controllers that can mitigate financial and reputational risk, then they are likely to be seen as such when personal data is being repurposed in a big data environment. And to the extent that the assessment is framed in broader terms than data protection, then the larger issues related to discrimination and social sorting may then be addressed.

However, there are also some extraordinary challenges. First, how does an organization assess the expected benefits when the analytical process is essentially an inductive "fishing-expedition" within the data? Big data analytics tend to be premised upon very vague assertions about the rewards to society, business and consumers. And whereas privacy professionals now have a familiar set of tools for assessing privacy risk, it is not clear how they assess and prioritize a project's potential rewards when the benefits are often so speculative (Polonetsky and Tene 2013). Further, how can PIAs be conducted when multiple organizations and data sources may be involved in a big data project, and where lines of accountability may become very blurred? PIAs should ideally be

transparent. Yet big data analytics often rely on the application of secret and proprietary algorithms, the understanding and assessment of which is necessary for the overall consideration of privacy risk. Thus PIAs are not sufficient to address the broader set of risks in a big data environment. In this light, two further sets of tools have been proposed.

Big Data and Surveillance Impact Assessments

Charles Raab and David Wright (2012) have introduced the concept of the "Surveillance Impact Assessment" (SIA) to respond to the critique that PIAs are too narrowly focused on individual privacy. In common with current usage, they adopt a quite broad definition of surveillance to embrace the systematic capture of personal data beyond that collected through visual means. In addition to "watching," surveillance is conducted through listening, locating, detecting, dataveillance, as well as through the combination of those practices in "assemblages" (2012, pp. 370-372). Raab and Wright conceptualize four nested concentric circles of an SIA. In the innermost circle (PIA1) falls the conventional PIA, focused on individual privacy. PIA2 adds other impacts on an individual's relationships, positions and freedoms. The third stage (PIA3) adds the impact on groups and categories. The fourth (outermost) ring of the circle (PIA4), adds the broader impacts on society and the political system. The model is intended to be cumulative. Under this framework, privacy progressively assumes the character of a social or collective good (Regan, 1992) as one extends the analysis to the outer limits of the circle.

How could this framework assist with the analysis of big data analytics? A concrete illustration might assist. Take the example of the practice of credit-scoring, based on neighborhood characteristics – crime rates, property values, and so on (World Privacy Forum, 2014). The impact of such a score on the individual's ability to get a loan would be regulated under the Fair Credit Reporting Act (FRCA) in the United States, but it would only apply to those aspects of the scoring system that were affected by an individual's credit report. This law, like other data protection acts, regulates the sources of legitimate personal data, and offers limited recourse to access and correct erroneous reports. A PIA1 would be thus confined to ensuring that the provisions of the law are appropriately considered when the credit-scoring system was put in place. At the next stage (PIA2), the impact of your credit score on your immediate social network would be analyzed. That could include neighbors in an immediate geographical sense, but also friends, family, and other individuals with whom one regularly associates, online or offline. At the third stage (PIA3), the analysis would expand to the effect of categories of individuals and groups. If you are the kind of person with a bad credit score, then inferences might be drawn about the credit-worthiness of similar people with a similar profile. And at the final stage (PIA4), one would consider the general workings of society: social and community relations, democratic rights, political participation, the impact on the criminal justice system and so on.

Wright, Friedewald and Gellert (2015) followed up this analysis with an attempt to develop a more explicit SIA methodology and tested it on four separate "smart" surveillance systems. Their project organized a series of scenario-based workshops to

which different stakeholders were invited to inform the project (SAPIENT) about the different drivers for surveillance technologies, the current legal rules on transparency and consent, the relative vulnerability of individuals, the possibilities of resistance and the variety of potential solutions. These authors also point out the limitations of existing PIA methodologies, and seek therefore a methodology that addresses wider privacy rights than just data protection, as well as other fundamental human rights and ethical values (p. 50). They too are insistent that the SIA should be conceived as a process, culminating in a published report that documents that process.

They then outline a twenty-step process in three phases (p. 51)

Phase I: Preparation
1. Determine whether an SIA is necessary.
2. Develop terms of reference for the surveillance assessment team.
3. Prepare a scoping report. (What is the scope of the surveillance system?)
4. Check compliance with legislation.
5. Identify key stakeholders.

Phase II: Risk identification and analysis
6. Initiate stakeholder consultation.
7. Identify risk criteria.
8. Identify primary assets and feared events. (What could happen if the surveillance system is implemented?)
9. Analyze the scope of feared events.
10. Analyze the impact of feared events.
11. Identify supporting assets.
12. Identify threats and analyze vulnerabilities.
13. Identify threat sources and analyze capabilities.
14. Create a risk map (for prioritizing risks for treatment).

Phase III: Risk treatment and recommendations
15. Risk treatment identification and planning
16. Prepare an SIA report.
17. Record the implementation of the report's recommendations.
18. Publish the SIA report.
19. Audit the SIA.
20. If necessary, update the SIA.

In the interests in keeping the process relatively simple, this methodology is obviously framed in quite high-level principles. They are also aware that overly complex and lengthy guidelines can scare away potential users. It is also obvious that once the assessment goes beyond legal compliance, then more subjective evaluations of risk inevitably enter the analysis. And those subjective assessments require consultations with a wide range of stakeholders, including the general public. It follows that the assessment has to explain the technology and the practice in a way understandable to the layman.

Both SIA methodologies reviewed here are probably more geared to public sector surveillance systems, and to projects, which have, some defined technical and institutional parameters. Much of the appeal of big data analytics is that there are often no parameters. As noted above, organizations are invited to "fish around" in the data until they find interesting correlations. Often there is no "project" as such, and thus no obvious starting point at which an SIA could begin, and an end-point at which it could be published. These realities only reinforce the broad conclusion that any risk assessments should be conceived as a process rather than an end-product.

Big Data, Accountable Privacy Governance and Ethical Impact Assessments

Under the auspices of the Information Accountability Foundation (IAF), certain privacy experts from the private sector have also been thinking about how to reconcile the promise of big data analysis with traditional privacy principles. The project is a work in progress (Information Accountability Foundation, 2014; 2015), but is worthy of comment. Unlike PIAs and SIAs, these assessments, it should be stated at the outset, are explicitly motivated by the question of how to analyze big data ethically.

This project builds upon earlier work by Martin Abrams and his colleagues to generate methods to encourage and measure organizational accountability (Center for Information Policy Leadership, 2009). Decision-making about the ethics of big data analysis is, therefore, inextricably connected to whether or not the organization has an effective privacy management framework in place. Logically, if overall privacy governance is done well within a company (and the project is focused mainly on the private sector), then it will have the systems (management and technical) in place to assess the risks and put the necessary safeguards in place. In Canada, the privacy commissioners have offered explicit advice about the various elements of good privacy governance (OPC, 2012). Thus, many of the ways that organizations might mitigate risk is to ensure that privacy is an integral part of an organization's commitments and governance structure.

The IAF's project on big data conceives of four integrated steps: A Unified Ethical Frame; An Interrogation Framework; An Enforcement Discussion; and Industry Interrogation Models. The project has begun with an analysis of the larger ethical considerations and will progressively drill down to more practical guidance for industry.

Part A of this project conceives a "Unified Ethical Frame" designed to ensure a "balanced ethical approach to big data" (IAF, 2015, p. 7). The paper identifies five core values: beneficial, progressive, sustainable, respectful and fair. The paper suggests an ethical review that goes way beyond data protection, privacy and surveillance. The ambition is to produce a framework that also embraces the individual rights and interests addressed in common declarations of fundamental rights such as the United Nations Charter of Fundamental Rights, including values such as health, education, and opportunity and benefits from technology, which are advanced by the data processing. It is also intended to encompass corporate interests in innovation and in return on investment.

These values then inform the Interrogation Framework designed to be used when "big data projects reach key milestones and decision points" (IAF, 2015, p. 3), meaning at the concept, discovery, application and review phases. The framework yields a worksheet designed to alert organizations to the key questions that need to be addressed at each stage to determine if the project is beneficial, progressive, sustainable, respectful and fair. This is explicitly an "Interrogation framework" rather than a more rigid set of guidelines. It is designed as a prompt that might be adapted for different companies and for different purposes. The key is that organizations have to be accountable and that they are able to demonstrate to regulators "that they have, effectively and with integrity, identified the full range of individual interests, and balanced those interests with other societal concerns" (2015, p.5). Clearly, the value of these tools will only be properly judged, when these higher-level instruments are applied to more specific industry applications.

Conclusions

Aside from technical solutions, is there a way to reconcile big data analytics with traditional information privacy principles? And what role might PIAs play in that process? We conclude with two separate attempts to grapple with these questions.

One approach is offered by Tene and Polonetsky (2013). These authors regard the information privacy principles not as a rigid framework, but as a "set of levers that must be adjusted to adapt to varying business and technological conditions. Indeed, the ingenuity of the FIPPs is manifest in their flexibility, which has made them resilient to momentous change—some principles retract while others expand depending on the circumstances. In the context of big data this means relaxing data minimization and consent requirements while emphasizing transparency, access, and accuracy" (p. 242).

They propose a set of solutions, which de-emphasize the role of individuals at the point at which personal data is captured. They concede the fundamental weaknesses of a notification and consent model that relies on opting into, or out of, data processing practices based on the non-reading of complex and legalistic privacy policies. Rather, they want to shift emphasis to the empowerment of individuals, which allow them to engage with the benefits of big data for their own particular usage. As a "quid pro quo for looser data collection and minimization restrictions, organizations should be prepared to share the wealth created by individuals' data with those individuals." Individuals would then have access to "their data" and would be able to make more useful choices about how to live their lives. They contend that the "featurization" of big data could unleash more applications and create a market for such end-user innovations. A model would be the smart-grid applications designed to allow users to monitor their energy usage and make more intelligent decisions about their appliance usage, and about energy consumption.

They concede that this call for greater transparency is not new. It is just that the mechanisms (requirements for the transparent notification of purposes, and individual access and correction rights) have not succeeded as regulatory tools. The entire "app

economy" is now, however, premised on individuals being able to access their own personal data to make intelligent choices about consumption, finance, health, and so on. It can, and should, be leveraged to provide individuals with access to their data in "usable" format and thus render the big data "ecosystem" more transparent. To this end, they propose that organizations not only reveal the existence of their databases, but also the criteria (not necessarily the algorithms) used in their decision-making processes, subject to protection of trade secrets and intellectual property. In this way, individuals could not only scrutinize the accuracy of the data, but also the reasonableness of the inferences drawn from that data (2013, pp. 270-71). For Tene and Polonetsky, the problem is not "big data" per se, but "secret big data."

A second approach is offered by the Center for Information Policy Leadership (2013), which formed the basis for the development of the IAF's Unified Ethical Framework, cited above. This 2013 paper sought to explain in a little more detail how big data analysis is actually conducted with a view to offering practical and effective privacy guidance. The paper insists that there is a crucial distinction between knowledge discovery and application. The former comprises acquisition, pre-processing, integration, analysis and interpretation. In each of these phases, algorithms perform a variety of classificatory, associational and sequential tasks (p. 10). It is only in the application phase, they argue, that insights about individuals might be enabled. For the most part, the knowledge discovery phase "does not involve analysis of a particular individual's data (which may be de-identified or pseudonymised) and does not result in decisions about him or her" (p. 14). Because the individual is implicated, but not affected, by the first phase, different protections are warranted.

Thus, privacy rules for big data need to: 1) recognize and reflect the two-phased nature of analytic processes; 2) provide guidance for companies about how to establish that their use of data for knowledge discovery is a "legitimate business purpose" (under the GDPR); 3) rely upon fair information principles but apply them in a manner appropriate to the processing of big data for analytics; 4) emphasize the need to establish accountability through an internal privacy programme; 5) take into account that analytics may be an iterative process using data from a variety of sources; 6) reinforce the importance of appropriate data security measures; and 7) foster interoperability across diverse jurisdictions. Like Tene and Polonetsky, the fair information principles are regarded as "a cornerstone for guidance" rather than a rigid set of regulatory requirements. In that light, notions of consent for the collection of data can, and should, be flexible, and assessed in the light of they ways that those data are used.

Are such distinctions possible, however? Many legal regimes have abandoned the attempt to distinguish between "collection" and "use." Those separate steps were a feature of regulation in the 1980s, and are expressed as separate principles in the OECD Guidelines of 1981 (OECD, 1981). However, it is not a distinction that features prominently in contemporary European law, which has favored one undifferentiated concept of "data processing" (EU, 1995). Furthermore, the emphasis on individual control as *the* central tenet of data protection law may also be misplaced. In many countries outside the United States, privacy regulation is underpinned by the assumption

that personal data processing requires more proactive oversight through specialized data protection authorities (DPAs). That recognition goes back to the beginning of the modern data protection movement (Simitis, 1978).

In conclusion, we agree with Kerr and Earle when they conclude that "the nexus between big data and privacy is not a simple story about how to tweak existing data protection regimes in order to 'make ends meet'; big data raises a number of foundational issues" (Kerr and Earle, 2013). But the model has always been under stress, and yet has been able to adapt and embrace the regulation and management of an enormous range of new technologies and practices. As the UK ICO concludes: "The basic data protection principles…are still fit for purpose…Big data is not a game that is played under different rules" (ICO, 2014b, p. 41).

Moreover, it is also crucial to regard the governance of privacy as embracing a package of different regulatory, self-regulatory, and technological policy instruments (Bennett and Raab, 2006), with both proactive and reactive elements. In this light, the development and application of broader, surveillance and ethical assessment tools can obviously play a central role in ensuring that big data analysis is conducted with appropriate regard for privacy and other values.

However, data controllers and data regulators also need to pay heed to prior advice (e.g. Warren et al. 2008; Wright and de Hert, 2012) about how PIAs should be conducted within existing privacy regimes. Privacy assessment tools need to: genuinely conduct a prospective identification of privacy risks *before* the data is analyzed and involving all relevant employees and consulting with key stakeholders; assess the impacts in terms *broader* than those of legal compliance; be *process* rather than output oriented; and use a *systematic* methodology. The challenges posed by these new analytical processes are real, to be sure. But organizations are less likely to face legal, financial, and reputational damage if they seriously heed existing advice about how accountable organizations should identify and mitigate risks and implement effective privacy management within their organizations.

REFERENCES

Bennett, Colin J. 1992. *Regulating Privacy: Data Protection and Public Policy in Europe and the United States,* Ithaca: Cornell University Press.

_____ 2013. "The politics of privacy and the privacy of politics: parties, elections and voter surveillance in Western democracies," *First Monday* vol. 18, no. 8 (August 5, 2013) at: http://firstmonday.org/ojs/index.php/fm/article/view/4789

Bennett, Colin J. and Charles D. Raab, 2006. *The Governance of Privacy: Policy Instruments in Global Perspective.* Cambridge: MIT Press.

boyd, danah & Kate Crawford. 2012. "Critical Questions for Big Data," *Information, Communication and Society* vol. 15: 662-667

Center for Information Policy Leadership. 2009. "Data Protection Accountability: The Essential Elements," October 2009 at: http://www.huntonfiles.com/files/webupload/CIPL_Galway_Accountability_Paper.pdf

Center for Information Policy Leadership. 2013. "big data and Analytics: Seeking Foundations for Effective Privacy Guidance" (Discussion Document, February 2013) at: https://www.hunton.com/files/Uploads/Documents/News_files/big_Data_and_Analytics_February_2013.pdf

Electronic Information Center (EPIC), 2014. Comments to the Office of Science and Technology Policy on "Big data and the Future of Privacy" April 4, 2014 at: https://epic.org/privacy/big-data/

European Union. 1995. *Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of Such Data.* Brussels: OJ No. L281, 24 October 1995.

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on General Data Protection Regulation (OJEU L119 1), available at <http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

Gandy, Oscar H. 1993. *The Panoptic Sort: A Political Economy of Personal Information* Boulder: Westview.

Hoofnagle, Christopher. 2014. "The Potemkinism of Privacy Pragmatism," *Slate*, September 2, 2014 at: http://www.slate.com/articles/technology/future_tense/2014/09/data_use_regulation_the_libertarian_push_behind_a_new_take_on_privacy.2.html

Information Accountability Foundation. 2014. "A Unified Ethical Frame for big data Analysis" (Draft, October 8, 2014) at: http://informationaccountability.org/wp-content/uploads/IAF-Unified-Ethical-Frame-v1-08-October-2014.pdf

Information Accountability Foundation. 2015. "IAF big data Ethics Initiative Interrogation Framework," (Draft March 18, 2015) at: http://informationaccountability.org/wp-content/uploads/IAF-big-Data-Ethics-Initiative-Draft-Part-B-Final-03-03-2015.pdf

_____. 2014. *Big Data and Data Protection*. Wilmslow: ICO at: https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

Kerr, Ian and Jessica Earle. 2013. "Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy," 66 *Stanford Law Review* at: http://www.stanfordlawreview.org/online/privacy-and-big-data/prediction-preemption-presumption

Lyon, David. 2003. *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination* London: Routledge.

Mayer-Schönberger and Kenneth Cukier, 2013. *Big Data: A Revolution that will Transform how we Live, Work and Think* New York: Houghton, Mifflin, Harcourt.

Mundie, Craig. 2014. "Privacy Pragmatism: Focus on Data Use, Not on Data Collection," *Foreign Affairs*, March/April 2014

National Consumer Law Center, 2014. *Big Data: A Big Disappointment for Scoring Consumer Credit Risk,* Boston: NCLC at: http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf

The New Transparency Project, 2014. *Transparent Lives: Surveillance in Canada*. Athabasca: Athabasca University Press.

Office of the Privacy Commissioner Canada (OPC), Office of the Information and Privacy Commissioner of Alberta, Office of the Information and Privacy Commissioner of BC. 2012. *Getting Accountability Right with a Privacy Management Program* (April 2012) at: https://www.priv.gc.ca/information/guide/2012/gl_acc_201204_e.asp

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* vol. 57: *1701.*

Organization for Economic Cooperation and Development (OECD), *Guidelines for the Protection of Personal Information and Transborder Data Flows*, (Paris: OECD, 1981)

at:
http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Polonetsky, Jules and Omar Tene. 2013. "Privacy and Big Data: Making Ends Meet," 66 *Stanford Law Review Online* 25 at: http://www.stanfordlawreview.org/online/privacy-and-big-data/privacy-and-big-data

Raab, Charles D. and David Wright. 2012. "Surveillance: Extending the Limits of Privacy Impact Assessment," in David Wright and Paul de Hert, *Privacy Impact Assessments*, Dordrecht: Springer.

Regan, Priscilla. 1995. *Legislating Privacy: Technology, Social Values and Public Policy*. Chapel Hill: University of North Carolina Press.

Robinson, David and Harlan Yu. 2014. *Civil Rights, Big Data and your Algorithmic Future*, Leadership Conference on Civil and Human Rights at: https://bigdata.fairness.io/wp-content/uploads/2014/11/Civil_Rights_big_Data_and_Our_Algorithmic-Future_v1.1.pdf

Schwartz, Paul M. and Daniel J. Solove, 2011. "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", 86 N.Y.U. L.Q. Rev. 1814: Available at: http://scholarship.law.berkeley.edu/facpubs/1638

Simitis, Spiros. 1978. Reviewing Privacy in the Information Society. *University of Pennsylvania Law Review*. Vol. 135: 707-746.

Stalder, Felix. 2002. Privacy is not the Antidote to Surveillance. *Surveillance and Society.* Vol.1: 120-124.

Tene, Omar and Jules Polonetsky. 2013. "Big Data for All: Privacy and User Control in the Age of Analytics," *Northwestern Journal of Technology and Intellectual Property*," Vol 11, No. 5: 239-273

Taylor, Scott. 2014. Conference highlights, International Data Protection Commissioner Conference, Mauritius. Report at: http://www.privacylaws.com/Int_enews_October14_4

United States, Executive Office of the President, 2014. *Big Data: Seizing Opportunities, Preserving Values* (The White House, May 2014).

United States President's Council of Advisors on Science & Technology. 2014. *big data and Privacy: A Technological Perspective*, The White House, May 1, 2014, at: www.whitehouse.gov/bigdata

Warren, Adam, Colin J. Bennett, Roger Clarke, Andrew Charlesworth and Robin M. Bayley 2008. "Privacy Impact Assessments: International Experience as a Basis for UK Guidance," *Computer Law and Security Report* vol. 24: 233-242.

Weitzner, Daniel J. et al. "Information Accountability," 2008. *Communications of the ACM,* June 2008: Vol. 51. No. 6: 84.

World Privacy Forum. 2014. *The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future* at http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf

Wright, David and Paul de Hert (eds). 2012. *Privacy Impact Assessments.* Dordrecht: Springer.

Wright, David, Michael Friedewald and Raphael Gellert, 2015. "Developing and Testing a surveillance assessment impact methodology," *International Data Privacy Law*, 2015, Vol. 5, No. 1.